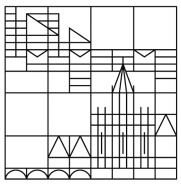
Longitudinal Research in Human-Computer Interaction

Dissertation zur Erlangung des
akademischen Grades eines Doktors der **Naturwissenschaften** (Dr. rer. nat.)
im Fach **Informationswissenschaft**

vorgelegt von
Jens Gerken

an der

Universität Konstanz



Mathematisch-Naturwissenschaftliche Sektion

Fachbereich Informatik & Informationswissenschaft

Tag der mündlichen Prüfung: 22.11.2011

Referent: Prof. Dr. Harald Reiterer

Referent: Prof. Dr. Kasper Hornbæk

Danksagung

Ich möchte meinen Eltern danken, die mich bei all meinen Entscheidungen hinsichtlich Studium und Promotion immer und in jeglicher Form unterstützt haben. Meiner Freundin Anja möchte ich besonders danken, da sie mich gerade in der „heißen“ Phase ertragen musste und immer für mich da war. Des Weiteren gilt mein besonderer Dank Prof. Dr. Harald Reiterer, der mir bereits in „frühen“ Jahren sehr viel Vertrauen geschenkt hat, mich auf Konferenzen Vorträge halten ließ und zu EU Projekttreffen geschickt hat. Darüber hinaus hat er mich immer unterstützt und gefördert. Auch bei den wichtigen Entscheidungen hinsichtlich beruflicher Perspektive war er mir ein wichtiger Gesprächspartner und Mentor. Von all den Kommilitonen und Kollegen über die Jahre, stechen zwei besonders hervor: Werner König und Hans-Christian Jetter. Gemeinsam haben wir das Studium geschafft, etliche Projekte durchgeführt, Paper geschrieben, Diskussionen über HCI, Gott und die Welt geführt und uns gegenseitig immer weiter voran gebracht. Hätten wir uns nicht so perfekt ergänzt, wäre wohl keiner von uns so weit gekommen. Stefan Dierdorf und Patric Schmid möchte ich für ihren großen Einsatz für das Projekt PocketBee danken. Ohne die beiden gäbe es heute kein lauffähiges System und keine erfolgreiche Industrie Kooperation. Ebenso möchte ich Thorsten Büring, Mathias Heilig, Jo Bieg, Michael Zöllner, Svenja Leifert, Roman Rädle, Daniel Klinkhammer und Alexandra Sautner danken. Mit allen habe ich tolle gemeinsame Projekte durchführen können und alle sind zu guten Freunden geworden. Natasa Milic-Frayling möchte ich für die tolle Zeit bei Microsoft Research in Cambridge danken. Abschließend gilt mein Dank der gesamten Arbeitsgruppe Mensch-Computer Interaktion, die dafür verantwortlich ist, dass ich mich dort über so lange Zeit so wohl gefühlt habe.

Abstract

The goal of this thesis is to shed more light into an area of empirical research, which has only drawn minor interest in the field of Human-Computer Interaction so far – Longitudinal Research. This is insofar surprising, as Longitudinal Research provides the exceptional advantage compared to cross-sectional research of being able to analyze change processes. Therefore, it incorporates time as a dependent variable into the research design by gathering data from multiple points in time. Change processes are not just an additional research area but are essential to our understanding of the world, with HCI being no exception. Only Longitudinal Research allows us to validate our assumptions over time. For example, a user experience study for an electronic consumer product, such as a TV-set, that reveals how excited people about the device are, should also investigate whether this excitement holds over time, whether usability issues arise after two weeks, and eventually whether people will buy the follow-up product form the same company. Our experience with technology is situated in context, and time is one important aspect of our context – ignoring does not necessarily lead to invalid but often insignificant research.

In this thesis, we contribute to the area of Longitudinal Research in HCI in manifold ways. First, we present a taxonomy for Longitudinal Research, which provides a foundation for the development of the field. It may serve both as a basis for discussion and methodological advances as well as a guiding framework for novices who strive to apply Longitudinal Research methods.

Second, we provide a practical contribution by presenting PocketBee, a multimodal diary for longitudinal field research. The tool is based on Android smartphones and allows researchers to conduct remote longitudinal studies in a variety of ways. We embed the discussion of PocketBee in a broader discussion of the diary and experience sampling methods, allowing researchers to understand the context of the tool, the advantages and also the inherent problems.

Eventually, we present the Concept Maps method, which tackles a specific issue of Longitudinal Research – the difficulty to analyze changes in qualitative

data over time, as these are normally hidden in large amounts of data and subject to the interpretation of the researcher. In the context of API usability, the method allows the externalization of the mental model developers generate. Concept Maps are used for these external representations and by continually updating these maps, changes over time become apparent and the analysis replicable.

The thesis will also help researchers to discover further important research areas in this field, as for example the variety of methodological issues that arise with gathering data over time. As the topic of Longitudinal Research has not yet been covered comprehensively in the scientific HCI literature, this thesis provides an important first step.

Zusammenfassung

Ziel dieser Arbeit ist es, mehr Licht auf einen Bereich empirischer Forschung zu werfen, welcher bislang innerhalb der Disziplin Mensch-Computer Interaktion nur geringe Beachtung fand: Längsschnittforschung. Dies ist insofern überraschend, da Längsschnittforschung gegenüber klassischer Querschnittforschung den entscheidenden Vorteil hat, dass Veränderungen über die Zeit analysiert werden können. Um dies zu erreichen, wird die Variable *Zeit* explizit in das Forschungsdesign integriert, indem Daten zu bzw. für mehrere Zeitpunkte erhoben werden. Dabei ist zu beachten, dass Veränderungsprozesse nicht nur eine weitere Forschungsmöglichkeit darstellen, sondern ganz entscheidend für unser Verständnis unserer Welt sind – und hier ist die MCI keine Ausnahme.

Längsschnittforschung ist als einzige in der Lage, unsere Annahmen was zeitliche Veränderungen zu betrifft zu validieren. Beispielsweise sollte eine User Experience Studie eines elektronischen Konsumproduktes (z.B. ein Fernsehgerät), die aufzeigt wie begeistert die Nutzer von dem Gerät sind, ebenso untersuchen, ob diese Begeisterung über die Zeit bestehen bleibt, ob sich Usability Probleme nach und nach zeigen und ob die Nutzer letztlich ein weiteres Gerät der gleichen Marke kaufen. Die Erlebnisse und Erfahrungen, die Menschen im Umgang mit Technologie sammeln, sind immer eingebettet in den Kontext der Nutzung und für diesen spielt Zeit eine entscheidende Rolle – diesen Faktor zu ignorieren führt nicht zwangsläufig zu fehlerhaften Ergebnissen, aber oftmals zu letztlich unbedeutender Forschung.

Diese Arbeit leistet auf mehreren Ebenen einen Beitrag zu dem Gebiet der Längsschnittforschung in der MCI. Zunächst wird eine Taxonomie für Längsschnittforschung vorgestellt, welche eine Grundlage für die weitere Entwicklung dieses Forschungszweiges darstellt. Sie kann hierbei sowohl als eine Ausgangsbasis für wissenschaftlichen Diskurs und methodische Weiterentwicklungen dienen als auch Interessierten, die tiefer in die Thematik einsteigen möchten, ein hilfreiches Framework sein.

Zum zweiten wird ein multi-modales Tagebuchwerkzeug, PocketBee vorgestellt und hiermit ein praktischer Beitrag für das Feld getätigt. Das Werkzeug basiert auf Android Smartphones und erlaubt es Forschern, Remote-Studien im Längsschnitt auf vielfältige Weise durchzuführen. Dabei wird PocketBee eingebettet in eine umfangreiche Diskussion von Tagebuch und Experience Sampling Methodik vorgestellt und diskutiert, wodurch interessierte Forscher die Möglichkeit erhalten, den Kontext des Werkzeuges, die Vorteile und auch die Prinzipbedingten Nachteile besser zu verstehen.

Schlussendlich adressiert die Concept Maps Methode einen spezifische Herausforderung von Längsschnittforschung: die Analyse von Veränderungen in qualitativen Daten. Diese sind zumeist in großen Datenmengen versteckt und unterstehen der subjektiven Interpretation des Forschers. Im Kontext von API Usability erlaubt die Concept Maps Methode die Externalisierung des mentalen Models, welches die Entwickler im Umgang mit der API gebildet haben. Dies geschieht über Konzeptkarten, welche zudem kontinuierlich erweitert und modifiziert werden. Hierdurch werden Veränderungen über die Zeit offensichtlich und die Analyse dieser wird nachvollziehbar und replizierbar.

Diese Arbeit soll auch dazu dienen, auf weitere wichtige Forschungsfelder aufmerksam zu machen, die durch die Datenerhebung über die Zeit zu Tage treten. Da die Thematik der Längsschnittforschung bislang in der Mensch-Computer Interaktion nicht umfänglich betrachtet wurde, stellt diese Arbeit hierzu einen entscheidenden ersten Schritt dar.

Parts of this thesis were published in:

Gerken, J, Bak, P & Reiterer, H 2007, 'Longitudinal Evaluation Methods in Human-Computer Studies and Visual Analytics', *Metrics for the Evaluation of Visual Analytics (InfoVis 2007 Workshop)*.

Gerken, J, Bak, P, Jetter, H-C, Klinkhammer, D & Reiterer, H 2008, 'How to use interaction logs effectively for usability evaluation', *BELIV 2008: Beyond time and errors (A CHI 2008 Workshop)*, ACM Press.

Gerken, J, Demarmels, M, Dierdorf, S & Reiterer, H 2008, 'HyperScatter – Modellierungs- und Zoomtechniken für Punktdiagramme', *Mensch & Computer 2008: Viel mehr Interaktion, 8. Konferenz für interaktive und kooperative Medien*, Oldenbourg Verlag.

Gerken, J, Heilig, M, Jetter, H-C, Rexhausen, S, Demarmels, M, König, WA & Reiterer, H 2009, 'Lessons Learned from the Design and Evaluation of Visual Information Seeking Systems', *International Journal on Digital Libraries*, August 2009, pp. 49-66.

Gerken, J, Bieg, H-J, Dierdorf, S & Reiterer, H 2009, 'Enhancing Input Device Evaluation: Longitudinal Approaches', *CHI 2009: Extended Abstracts*, ACM Press.

Gerken, J & Reiterer, H 2009, 'Eine Taxonomie für Längsschnittstudien in der MCI', *Mensch & Computer 2009*, Oldenbourg Verlag.

Gerken, J, Dierdorf, S, Schmid, P, Sautner, A & Reiterer, H 2010, 'PocketBee: a multi-modal diary for field research', *nordiCHI: In Proc. of the 6th Nordic Conference on Human-Computer Interaction*, ACM Press.

Gerken, J, Jetter, H-C & Reiterer, H 2010, 'Using Concept Maps to Evaluate the Usability of APIs', *CHI 2010: Extended Abstracts*, ACM Press.

Gerken, J, Jetter, H-C, Zöllner, M, Mader, M & Reiterer, H 2011, 'The Concept Maps Method as a Tool to Evaluate the Usability of APIs', *CHI'11: Proceedings of the 29th international conference on Human factors in computing systems*, ACM Press.

Table of Content

List of Tables

List of Figures

1 Introduction

Continuity gives us roots; change gives us branches, letting us stretch and grow and reach new heights. ~Pauline R. Kezer

Longitudinal research is often understood as synonymous with empirical research that lasts for a (very) long time, at least months or years. However, this is quite a limited perspective that focuses on the practical means rather than the design rationale; it does not capture the true value of and motive for doing longitudinal research. Rather, the driving principle behind such research is the realization that our world is highly dynamic. As research aims at understanding the world, we must take this into account. While observing the object of study for a prolonged period of time is certainly a key ingredient to longitudinal research, there is much more involved, as we will show in this thesis.

While this thesis will focus on longitudinal research in HCI, we take into account the perspective of adjacent disciplines from the social sciences and psychology as well. Especially within the social sciences, there is a long tradition of longitudinal research. Menard reports that what were probably the first systematically periodic censuses were conducted in New France and Quebec from 1665 to 1754 (Menard, 2002, p. 1). While censuses had been collected prior to this time (in ancient Rome, for example), the New World procedure allowed the systematic analysis of change processes, as the criteria for participant selection and the data gathered remained to some extent stable. As such, these censuses adopted a so-called repeated cross-sectional design. This form of longitudinal research is especially popular in survey research, as it basically combines several discrete cross-sectional studies, thereby reducing the organizational costs of tracking the same people across several years, as in longitudinal panel designs. On this subject, Menard cites several studies from as early as 1759 that examined individual change processes.

Technically, longitudinal research can best be described by contrasting it to cross-sectional research. In cross-sectional research, there is only one single measurement for each individual or each case in the study. In the best-case

scenario, the measurement for each individual and for each variable would happen at the same time, thus “regarded as contemporaneous” (Menard, 2002). In contrast, longitudinal research takes at least two measurements for each case and for the same variable, thereby providing the means of comparing data between or among time periods. The term “measurement” should be regarded in a rather broad sense here, including any kind of data-gathering. Longitudinal research is not a single method, but rather a set of methods or a research paradigm that is based on longitudinal data. Longitudinal data can 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. *(Taris, 2000, p. 1)*

While we will present a more elaborate classification of longitudinal research designs, all are based on three basic designs. Prospective longitudinal panel designs follow the same set of participants over multiple data-gathering sessions. Retrospective panel designs, on the other hand, query participants about multiple time points in the past, thereby asking them to recall certain events or feelings at these particular points in time. As mentioned, repeated cross-sectional designs use a different but comparable set of participants for each data-gathering session. In the social sciences, this is often not just a matter of convenience, but important to allow analysis of the influence of extra-individual change processes, such as changes in societies. For example, a longitudinal panel design could never be used to assess the research question of how adolescents perceive smoking today as compared to in the 1950s.

1.1 Why Do We Need Longitudinal Research?

But why should we actually consider conducting longitudinal research? As Menard states, longitudinal research in general incurs higher costs and has the same general problems as cross-sectional research, along with several addi-

tional issues that we will discuss later in detail (Menard, 2002, p. 78). However, it allows the study of research questions that simply cannot be answered through cross-sectional research. We will describe this type of research question in the next chapter in detail; in short, “pure” cross-sectional research does not tell us anything about intra-individual or inter-individual change processes. We cannot investigate how a person changes his or her opinion about a certain matter over time, for what reasons, or to what extent. We cannot study how the perceived quality of human relationships with and without children changes as time passes (Bleich, 1999). Market research could not investigate how the level of consumption of certain products changes over time, or whether two different products are recommended to the same extent directly after purchase and 6 months later. In addition to describing change processes, longitudinal research also allows us to see whether we can predict differences within these changes – for example, by comparing these two products.

Therefore, Menard’s conclusion is not surprising:

> The conclusion is inescapable, however, that for the description and analysis of dynamic change processes, longitudinal research is ultimately indispensable. *(Menard, 2002, p. 80)*

He continues:

> It is also the case that longitudinal research can, in principle, do much that cross-sectional research cannot but that there is little or nothing that cross-sectional research can, in principle, do that longitudinal research cannot. *(Menard, 2002, p. 80)*

However, he also stresses that longitudinal research is not the solution for everything. It cannot cancel out poor research design; on the contrary, it will probably magnify such mistakes or problems. It is also not necessary for every research question; it should be considered a tool for research, not as the ultimate or only method.

In Human-Computer Interaction, longitudinal research is still the exception to the rule, but it seems that during the last few years the need for such research

has consistently increased (Gerken, Bak, & Reiterer, 2007). This is exemplified by a growing number of activities at global conferences covering the topic. For example, the UPA conferences in 2005 and 2008 held a seminar (Gorlenko, 2005) and a workshop (Courage, Rosenbaum, & Jain, 2008) on this topic. In addition, several activities at the CHI conference exploring best practices of longitudinal research in academia and industry have been organized, including two special interest groups (Vaughan & Courage, 2007) (Jain, Rosenbaum, & Courage, 2010), a workshop (Courage, Jain, & Rosenbaum, 2009), and a panel discussion (Vaughan, et al., 2008)[1]. Additionally, several HCI researchers have explicitly stated the benefits that could be derived from such methods. Gonzáles and Kobsa (González & Kobsa, 2003), 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 (Saraiya, North, Lam, & Duca, 2006) 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. (Kjeldskov, Skov, & Stage, 2005) analyzed how the usability of a patient record system was perceived over time, concluding that "more longitudinal studies must be conducted into the usability of interactive systems over time, focusing on qualitative characteristics of usability problems." As early as 1999, MacKenzie and Zhang (MacKenzie & Zhang, 1999) stated that when comparing an 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."

Karapanos et al. argue that as products become more and more service-oriented, measurements of user experience will have to shift "from initial purchase to establishing prolonged use" (Karapanos, Zimmermann, Forlizzi, & Martens, 2010). Studies of User Experience (UX) are an intuitive example of the

[1] The results of these activities are available online: http://longitudinalusability.wikispaces.com

need for longitudinal research in HCI, as the concept of UX itself is not meant to remain stable over time and has shown to fluctuate significantly (Karapanos, Zimmermann, Forlizzi, & Martens, 2009). In their paper, Karapanos identified three different phases of product adoption: orientation, incorporation, and identification. Anticipation of use, both before purchase and before each use, affects these phases. During the orientation phase, users are most concerned with ease of use and stimulation. While excitement is often based on the discovery of novel features, frustration is related to learnability problems. The incorporation phase, in contrast, focuses on the usefulness of a product and its various features and whether the product has actually become a meaningful and significant part of the daily life of the user. The identification phase leads to an emotional attachment with the product as it is fully incorporated into daily life and even plays a role in social relationships with other users (Karapanos, Zimmermann, Forlizzi, & Martens, 2009).

Another important area for longitudinal research is the study of learnability. One of the first longitudinal studies in HCI, by Card et al. (Card, English, & Burr, 1978), compared several input devices with respect to their learnability. Also the study cited above by Kjeldskov et al. (Kjeldskov, Skov, & Stage, 2005) could be described as a learnability study, as it focused on whether usability issues disappear over time and could therefore be more precisely referred to as learnability issues. As learning is inherently time-dependent, only longitudinal research is capable of capturing this dynamic process. Recently, Grossman et al. (Grossman, Fitzmaurice, & Attar, 2009) conducted an extensive literature review on the concept of learnability. They cite different definitions from a variety of authors and conclude:

> The above definitions give indication that there is no agreed upon definition for learnability. Even those definitions which only apply to initial learning, base their definitions on differing assumptions about the user and what the important measures are (i.e. errors, usage time, etc.). *(Grossman, Fitzmaurice, & Attar, 2009)*

This survey of learnability studies revealed that most studies used the term learnability without any definition, while others referred to aspects such as *first*

time performance, *change in performance over time,* or *ability to master system.* Based on this, the authors defined a taxonomy for learnability (see Figure 1).

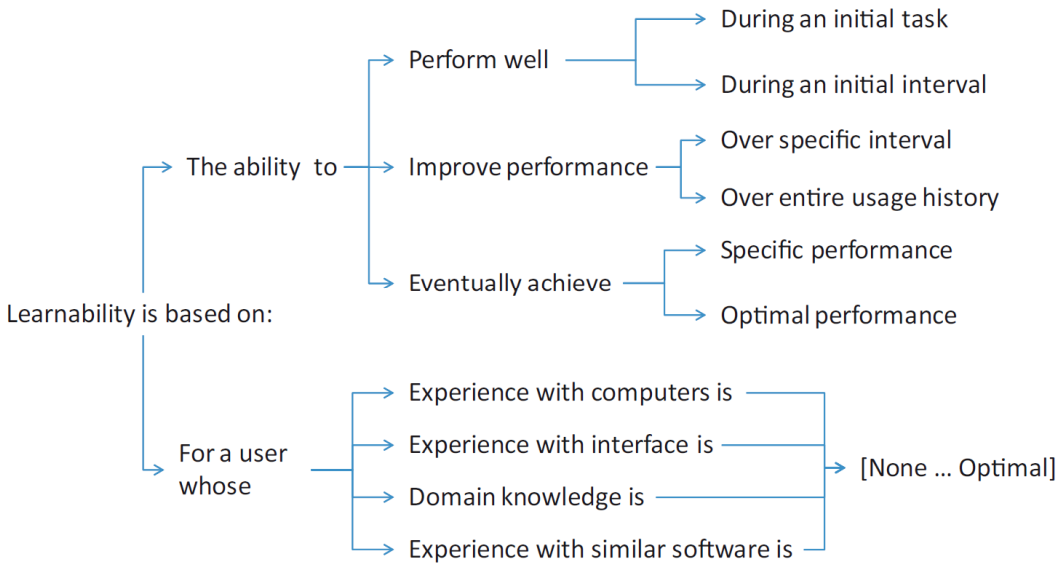


Figure 1: A taxonomy for learnability (Grossman, Fitzmaurice, & Attar, 2009)

The upper part of this figure presents various research questions in learnability. While the ability to perform well during an initial task/interval can in principle be studied in cross-sectional designs, all other aspects require a longitudinal design. Even for initial task performance, one could argue that only a longitudinal design would allow distinguishing between learnability and usability.

Unfortunately, Grossman et al. do not relate the study of learnability to longitudinal research, which, as we have discussed, we consider a necessary design for this type of research. What is also important to note is that the term “performance” should be interpreted broadly, covering not only typical performance measures such as task time or task effectiveness. Rather, overcoming learning barriers or the development of a correct mental model over time should also be considered here. In Chapter 4, we address these two aspects with our Concept Maps approach, which permits study of the usability and learnability of Application Programming Interfaces.

As stated above, longitudinal research is not a single method, but rather a research paradigm or a set of methods. In past applications of longitudinal research in HCI, the variety of approaches is apparent. For example, Saraiya et al. (Saraiya, North, Lam, & Duca, 2006) used as their methodological basis diaries in which insights and screenshots were stored by the participants themselves. Their goal was to get a better picture of the entire visual analytics process. On the other hand, Shneiderman and Plaisant (Shneiderman & Plaisant, 2006) present an approach designed as a field study that relies on many different data collection methods, such as interviews, observations, and logging. This is 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 the evaluation of information visualization and present descriptive guidelines for conducting such studies. MacKenzie and Zhang (MacKenzie & Zhang, 1999) rely on a series of laboratory-based studies to analyze how much training is necessary for a new soft-keyboard layout to become superior to the QWERTY standard. Kjeldskov et al. (Kjeldskov, Skov, & Stage, 2005) also rely on two laboratory studies to analyze whether usability problems could disappear after 15 months of system usage. One common aspect of most of these studies is the lack of explanation of why the specific longitudinal methodology was applied, making it difficult for other researchers to gain better understanding of the design space for longitudinal research. However, it becomes clear that this design space is much more complex than “doing a field study” or “studying something for a very long time.” In Chapter 2, we will shed light on this matter by presenting a taxonomy for longitudinal research in HCI.

1.2 Challenges in Longitudinal Research

Given all these advantages, one might assume that longitudinal research would be much more popular among researchers. However, longitudinal research is far from being without any obstacles or methodological challenges. In this section, we provide an overview of the most significant challenges. Some of them will be subsequently addressed more in detail in this thesis. As a basis for dis-

cussion, we will assume a longitudinal panel design, in which the same group of participants is followed over multiple data-gathering sessions. Other designs can address some of these challenges or introduce additional ones; this will be discussed in Chapter 2.2 when we present the taxonomy for longitudinal research designs. We can classify the challenges into organizational challenges and methodological challenges.

1.2.1 Organizational Challenges

The most predominant organizational challenge is the **cost factor**, and there is no denying that longitudinal research is in most cases more expensive than cross-sectional research. There are several different costs associated with this research design. First, a longitudinal study will generally take **more time to conduct**, as multiple data-gathering waves must be scheduled, prepared, analyzed, etc. Second, it is more **expensive to sample participants** for such a study, as participants must commit to a longer time period, a more complex study schedule, or both. Third, longitudinal study design and analysis is rarely part of researchers' education in HCI, therefore, **additional costs for advanced training** might be necessary. Eventually, as more time is committed, results are also delayed, which in an industrial context can be critical and result in **higher costs if the research fails**. However, these costs come with added value, as longitudinal research allows scientists to address completely new research questions. Any cost-comparison has to take this added value into account. Regarding the choice between cross-sectional and longitudinal research, Menard claims, "The choice should be between doing the research properly or not doing it at all" (Menard, 2002).

In addition to these cost factors, there are several more subtle organizational challenges. First, as a researcher, one must begin to **think in longitudinal research questions**. Simply extending a study over a prolonged time period, as we will discuss in the taxonomy of research questions in Chapter 2, will not automatically provide longitudinal benefits. It may be more expensive to have participants take part in a longitudinal study, but monetary costs are only one of the issues. Many participants have difficulty anticipating the effort needed for partic-

ipation, leading to the problem of **panel attrition** – i.e., participants unexpectedly dropping out of the study. Therefore, a more thorough introduction to the study is necessary, as are incentives to keep motivation high. The **personal relationship** between researcher and participant also plays a much bigger role; establishing a comfortable situation can be critical to a study's success.

1.2.2 Methodological Challenges

In contrast to the organizational challenges, which one can either address (as with the increased effort needed for the relationship between researcher and participants) or simply accept (as with the higher costs), methodological challenges are not so easy to resolve. Most are inherent to longitudinal research or specific longitudinal designs; in most cases, they increase the difficulty of achieving valid results.

1.2.2.1 Panel Conditioning

According to Cantor, panel conditioning means that participants are conditioned (i.e., influenced) through participation in the study and their behavior in later data-gathering periods is thereby affected. The consequence is that "the result [of a study] is partly a function of the measurement process" (Cantor, 2008). Cantor cites Waterton and Lievesly (Waterton & Lievesley, 1989), who discussed several reasons for conditioning. For example, they found that raised consciousness in participants can result in changes in behavior or attitudes. Participants often try to figure out what the researcher wants to achieve. In essence, this means that participants begin to think about the subject of the study more and more and may adjust their behavior accordingly – often to fit what they think is expected by the researcher. An improved understanding of the study requirements (i.e., what the participants are meant to do, how they are supposed to understand certain questions, etc.) can influence participants and thereby introduce a bias. Increased or decreased motivation also introduces a bias that can confound the results, as participants may suddenly try harder or stop trying. As Sturgis et al. point out, one issue with panel conditioning is that most existing studies either fail to clarify the underlying mechanisms of the con-

ditioning effects, or they study panel conditioning using study designs that themselves confound the effect of conditioning. (Sturgis, Allum, & Brunton-Smith, 2009). Moreover, when not explicitly investigating panel conditioning, it is very difficult to assess the effect this process might have on results. Basically, one will never know for certain whether conditioning took place and to what extent.

There are only few methods to preemptively reduce the potential effect of panel conditioning. The most promising are revolving panel designs, which we will describe in Chapter 2; this strategy integrates a new set of participants at each data-gathering wave. When participants are not exposed to an experimental condition, this is a complex but achievable approach. If an experimental condition is introduced, then such a revolving panel can only reduce the conditioning effects introduced by the measurement or observation, but obviously not those of the experimental condition. Another way to avoid conditioning effects (while also introducing other problems) is utilization of a retrospective panel design, which we will also discuss extensively in Chapter 2.

Cantor presents a classification of various conditioning effects (Cantor, 2008):

- Changes in behavior caused by the data-gathering process: Cantor gives the example that people who have been interviewed about voting prior to an election are more likely to actually vote. Similarly, when we study technology adoption, we must ask critically whether the people we are studying are perhaps more likely to adopt technology simply because they are part of the study. Sung et al. (Sung, Christensen, & Grinter, 2009) report that some of their participants did not make any use of a house-cleaning robot that was provided as part of the study. They also stress that they undertook considerable effort to convince participants that they were absolutely free to use or not use the robot.
- Changes in report of behaviors, although participants have not actually changed: In many cases, it might be that participants do not actually change, but report their behaviors differently over time. Cantor cites several medical studies in which participants reported fewer medical issues over time. One possible reason is that participants might have tried to avoid the extra work

involved with taking part in the study. Cantor hypothesizes that in the case of an interview protocol that is repeated over time, participants begin to understand which answers lead to more questions (e.g., reporting changes or events), and so they try to avoid this extra effort. Another bias might arise if participants asked for changes get the feeling that they should have something to report, and thus start to make things up so that they might be considered “good” participants.

- Changing latent traits, such as attitudes, opinions, and subjective phenomena: Cantor reports that results for these types of variables are mixed, and that panel conditioning cannot be naturally assumed. One obvious example is when participants are asked to state an opinion about a certain matter they are not accustomed to considering; this may trigger them to actually inform themselves and form an opinion.

Cantor reports that effects of conditioning can be quite large: about 5-15% in effect size. However, it is unclear how dependent this size is on the research question and test instrument. He consequently concludes that much more research is required to get a better understanding of these effects and their influence on the validity of longitudinal data.

1.2.2.2 Construct Validity over Time

Another problem inherent to longitudinal research is that we cannot be sure that our measurement tool measures the same construct as time goes by. The problem is that “just because a measurement was valid on one occasion, it would not necessarily remain so on all subsequent occasions even when administered to the same individuals under the same conditions” (Singer & Willett, 2003, p. 14). This is certainly an issue for survey and questionnaire tools, and the problem goes beyond the conditioning effect described above (although it is a related effect). We will discuss this issue again in Chapter 2.3.8 with regard to data-gathering techniques and will focus here on two examples to illustrate the problem. The first is a classic example from educational research, as reported by Patterson (Patterson, 2008). When administering IQ tests over time from infancy to childhood, one cannot simply use the same test instrument, as infants

would not be able to “complete” the IQ test suitable for children, and using the infant test for children would no longer measure the older subjects’ IQs. The second example illustrates the possible relationship to panel conditioning. In a study by Mendoza and Novick (Mendoza & Novick, 2005), participants were asked to report frustrating episodes over the course of the study. However, what is experienced as “frustrating” may change over time. While the study seeks to investigate how frustration changes over time, the question remains of whether the construct itself is stable or changes due to the earlier frustrating experiences.

Again, there is no real solution to this issue other than varying the test instrument (as in the IQ study) or using a different longitudinal design (with its own shortcomings).

1.2.2.3 Panel Attrition

We have discussed panel attrition as an organizational issue; in this case, the focus is on ensuring that panel attrition is minimized. From a methodological point of view, panel attrition is also a severe problem. Menard points out several questions that one should ask in the case of panel attrition (Menard, 2002, pp. 39-40):

- Are those participants who left the panel different in a particular variable of interest compared to those who remain? If yes, to what extent and why?
- Is there a certain pattern of attrition, or is it random? In many cases, it will be time-dependent; i.e., as the study continues, a higher percentage will drop out. However, there might be a certain peak that requires further investigation.

Menard stresses that researchers should test their data for these questions and interpret their results accordingly. As an example, we refer back to the study by Mendoza and Novick (Mendoza & Novick, 2005). The authors state that 48 participants completed a pre-study questionnaire and that 32 of these provided reports for the full duration of the study. Let us assume that the other 16 provided

reports for some time but not over the complete duration.[2] If that were the case, Mendoza and Novick should check whether there is a certain pattern of frustration in the reports these "drop-outs" delivered, and whether they filled out more or less than the average participant. Let us assume that these 16 were much less active than the average participant from the beginning. There are at least two possible explanations: 1) they were not really motivated to participate in the study, explaining the low number of frustration episodes reported and the drop-out, or 2) they encountered only very few frustration episodes and at some point decided that taking part in the study was pointless, as they did not have anything to report. Without additional information, it is impossible to choose either one of these alternatives, but this decision has a tremendous influence on how to treat the data of these participants. In the first case, it might be acceptable to drop the participants completely and not consider their data in the overall analysis. However, in the second case, this decision would be harmful, as the remaining data would be biased towards more frustration episodes overall. Thus, even if panel attrition cannot be completely avoided, it is important to get as much data as possible about the drop-outs and their reasoning.

In addition to this important consideration, there is also a technical problem regarding data analysis. As we will discuss again in Chapter 2, one of the most commonly used statistical methods for data analysis, the analysis of variances (ANOVA) – or, in the case of a longitudinal study, a Repeated-Measures ANOVA – is unable to handle missing data. When data is missing, the researcher must discard data from drop-outs completely or use extrapolation, a potentially misleading and speculative technique that should only be used with great caution and for variables that are known to not change much. As we will see, there are other statistical methods, such as multi-level growth-curve modeling (Luke, 2008), that are more suitable here to allow incorporation of partial data into an analysis. Based on our literature review, it seems that these advanced statistical

[2] This is actually not apparent from the paper. It might very well be that Mendoza and Novick purposefully decided to leave out the 16 participants from the start, perhaps because they did not meet certain study requirements. We use this study only as an example to illustrate the issue.

methods are not yet common in HCI – which is not surprising, as Singer and Willett criticize the same issue for the social sciences (Singer & Willett, 2003). This refers back to one of the organizational challenges: Longitudinal research requires certain skills that are not yet common in HCI researchers, thus necessitating advanced training.

1.2.2.4 Data Analysis

We have already stressed this issue and will do so in the following chapters as well. Nevertheless, choosing an appropriate data analysis technique is important enough to merit its own section. For cross-sectional research, researchers are advised to pick the analysis technique before conducting the study; this is even more vital for longitudinal research. We see two reasons for this: First, in many cases the standard approaches are simply not appropriate. An experienced researcher in cross-sectional studies will know the tool box of methods that can be applied. When conducting one's first longitudinal study, one should not make the mistake of relying on previous experience; everything should be planned as well as possible in advance. The second reason is that for longitudinal research, data-gathering methods and analysis are much more interwoven with each other. The data-gathering needs to specifically address the change aspect and thereby dictates what kind of analysis is possible. This is an issue to a lesser extent with quantitative data, as long as certain aspects (such as the scheduling of data-gathering) are considered. For qualitative data, we find this to be absolutely essential. In Chapter 4, we will present the Concept Maps approach, which exemplifies how closely related data-gathering and analysis techniques in the case of qualitative longitudinal data can and should be.

Good advice for all varieties of longitudinal research (and also cross-sectional research) is provided by Singer and Willett:

> Wise researchers conduct descriptive exploratory analysis of their data before fitting statistical models. *(Singer & Willett, 2003, p. 16)*

1.3 Contributions

This thesis will contribute to the field of Longitudinal Research in HCI in a variety of ways, which we will briefly outline here. Based on our own experiences and the literature, we have identified three main challenges we would like to address. First, researchers in HCI lack a basic understanding of longitudinal research. As has been apparent from several workshops, SIGs, and panel discussions, there is neither a clear unifying view nor any basic literature to which people can refer. This makes it difficult to discuss issues in longitudinal research as well as to identify potential research areas that should be addressed. Second, especially for longitudinal field studies, we need more tools and techniques that could support researchers conducting such studies, thereby reducing the costs and any apprehension about getting involved in longitudinal research. Third, we need specific, tailor-made methods for longitudinal data-gathering and analysis, especially in the context of qualitative data.

1.3.1 A Taxonomy for Longitudinal Research in HCI

In Chapter 2 we will address the first issue, regarding the common understanding of longitudinal research in HCI. To this end, we will provide a theoretical workup of the topic that will eventually lead towards a taxonomy for longitudinal research in HCI. The goals of this taxonomy are 1) to give order to the existing literature in the field, taking into account findings from other disciplines, such as the social sciences and psychology; 2) to provide guidance for researchers and practitioners new to the field, helping them with an overview of the design space of longitudinal research; and 3) to promote scientific discussion by providing a common ground to which everyone can refer. The taxonomy is intentionally not restricted to a certain type of longitudinal research in HCI. Rather, by “going broad,” we would like to encourage other researchers to challenge the taxonomy, test it, and extend it, if necessary.

1.3.2 PocketBee - A Multi-modal Diary for Longitudinal Field Research

In Chapter 3, we will address the issue of tool support for longitudinal research. Based on the taxonomy presented in Chapter 2 we identify two areas that offer potential: interaction logging and diary/ESM approaches. As it is the more flexible tool, we opted for diary/ESM approaches. The chapter presents an exhaustive discussion of diary and ESM approaches, their advantages and drawbacks, before eventually leading to a discussion of PocketBee, a multi-modal diary tool based on Android smart phones. We contribute towards this field by presenting a classification of research designs that unifies diary and ESM studies and by providing a direct link to an event architecture that allows free combination of these designs within the PocketBee tool. Finally, we present the user interface design of the tool for participants and researchers, seeking to provide high usability and flexibility in methodology. In addition, PocketBee especially focuses on a closer connection between researcher and participant.

1.3.3 Concept Maps – A Method to Evaluate API Usability

In Chapter 4, we address the third issue by presenting a customized longitudinal data-gathering and analysis method for evaluating application programming interfaces (APIs). We present a constructive approach that implicitly asks participants to illustrate changes over time, allowing the researcher to easily identify them – an issue that can be very difficult with qualitative data. We focused on APIs because the issues of learnability and usability over time are of particular importance here. An API is not learned once and then applied; rather, programmers learn an API on the fly and to the extent needed for the task at hand. In addition, API usability is an often-overlooked aspect of overall product quality, which we found to be well worth additional consideration within the scope of this thesis.

2 A Taxonomy for Longitudinal Research in HCI

Any kind of empirical research needs to be designed. Even though the phrase "research method" conveys the idea of a clear step-by-step guide to solving a research question, this is hardly the case; such assumptions instead lead to uninspired and inappropriate research. Applying any kind of research paradigm requires the researcher to be aware of and acquainted with the design space the paradigm provides. Design space is a term often used in traditional design disciplines, such as graphical design or interaction design in HCI. The term refers to a space of possibilities for design within certain boundaries and featuring key attributes. Defining a design space basically means defining these boundaries and attributes. While HCI literature offers assistance in defining the design space for cross-sectional methods (including usability tests, experiments, interviews, and surveys), the research paradigm of longitudinal research clearly lacks such guidance. For instance, the major textbooks on research methods in HCI donate very little space to this topic (Rogers, Sharp, & Preece, 2007), (Cairns & Cox, 2008), (Lazar, Feng, & Hochheiser, 2009).

There are a number of different ways to describe the design space. One way that has attracted interest in HCI and software engineering is through patterns, which have also been used in interaction design. Design patterns provide examples that illustrate the basic principles of an applied design, how it was created, and whether it was successful. The patterns often try to incorporate these aspects into a single holistic and interlinked graphical representation (e.g. (Borchers, 2001)). However, obtaining an overview is often difficult (although not always necessary). Another possibility, which we address in this thesis, is the definition of a taxonomy. A taxonomy "refers to classification according to presumed natural relationships among types and their subtype"[3]. The major advantage we see in the taxonomy approach is its inherent structure and clarity that allows readers to quickly comprehend the entire design space without hav-

[3] ISO/IEC 11179, 1.

ing to understand all the specifics. Details are available but are confined to lower levels in the hierarchy of the taxonomy to the point which specifically asks for this kind of information.

While a practitioner should benefit directly from access to such a taxonomy, we think its value is much more extensive. As researchers seeking to take the methodologies of longitudinal research in HCI one step further in their development, we think it is essential to share a common overview of the current state. This allows us to identify the areas that need further research, no matter whether they concern new methods, new tools, or different theoretical understandings. For this thesis, the taxonomy has already served this purpose, as two major problematic areas of longitudinal research in HCI were successfully identified and subsequently addressed with the Concept Maps approach for API evaluation and the PocketBee diary/ESM tool (see chapter 3 & 4).

Longitudinal research in HCI is a very broad topic and it must seem to be a difficult task to define and carve out a general taxonomy. However, as longitudinal studies in HCI are still rare, we feel that limiting to a specific type of research area would be too restricting and leave too many areas uncovered. Therefore, our goal here is to provide the first step for a holistic taxonomy, being aware that we are likely to miss certain research areas. Our hope is that researchers of these areas will take the chance to build upon our taxonomy and extend or modify it, accordingly. To give the reader some perspective on our background, most of our own experiences with longitudinal research come from the domain of pointing device evaluation (Gerken, Bieg, Dierdorf, & Reiterer, 2009a), information visualization (Gerken, Demarmels, Dierdorf, & Reiterer, 2008b), and API usability (Gerken J. , Jetter, Zöllner, Mader, & Reiterer, 2011).

In the remainder of this chapter, we will present the taxonomy step by step. We will start by outlining the approach that eventually led to this taxonomy – a mixture of experiences gathered through the design of longitudinal studies, an extensive literature review of longitudinal research in other fields (including social sciences and psychology), and a review of HCI literature and in particular empirical studies that claim to be longitudinal.

2.1 The Approach

While taxonomy, as stated in the definition above, in its original sense is a classification of presumed natural relationships, we adapt a slightly different perspective here. As we are not modeling natural relationships but human defined research questions and designs our taxonomy seeks to integrate a more constructive perspective as well. This means that we consider the practical applications of longitudinal research as well as the boundaries for (statistical) analysis, which clearly must be taken into account in order achieve a taxonomy that does not describe research designs which have no practical outcome, or with data that cannot be analyzed.

The approach itself was not a linear or step-wise *production* of a taxonomy. Rather, many versions were created and abandoned along the way, as new knowledge had to be integrated. The basic ingredients, however, stayed the same. First, as other disciplines have much more experience with longitudinal research, such literature was taken as the primary data source. This included literature on both the design and the analysis of longitudinal research (e.g. (Menard, 2008), (Singer & Willett, 2003)) that seek to be multi-disciplinary, although without taking HCI research explicitly into account. Second, the author of this thesis designed and conducted several longitudinal studies in a variety of settings and with different research questions in mind (e.g. (Gerken, Demarmels, Dierdorf, & Reiterer, 2008b), (Rieger, 2009), (Gerken, Bieg, Dierdorf, & Reiterer, 2009a), (Gerken, Dierdorf, Schmid, Sautner, & Reiterer, 2010b), (Gerken J. , Jetter, Zöllner, Mader, & Reiterer, 2011)). The experience gained over the course of these studies was also incorporated into the construction of the taxonomy. Third, while there is no overview literature for longitudinal research in HCI, there are many published studies that claim to implement a longitudinal design. A literature review was conducted to analyze and categorize these studies. The requirements for the papers to be included in the review were a) they were published at major conferences (e.g., CHI) or in journals (e.g., International Journal of Human-Computer Studies) and b) that they reported the necessary details to replicate the research methodology. Based on these two requirements, a total of 42 papers could be included in the reviews,

spanning the period from 1978 to 2010 (see Appendix A). In addition, the author of this thesis participated in a variety of events concerned with longitudinal research in HCI, most notably three events that took place at the last three meetings of the International Conference on Human-Factors in Computing Systems (CHI). In 2008, a panel on longitudinal research was organized by Vaughan et al. (Vaughan, et al., 2008). In 2009, a full one-day workshop took place, organized by the same group of researchers (Courage, Jain, & Rosenbaum, 2009); the author presented a paper about this topic there (Gerken, Bieg, Dierdorf, & Reiterer, 2009a). In 2010, a SIG took place that featured a great deal of discussion about methods applied in longitudinal research but also the risks associated with the paradigm (Jain, Rosenbaum, & Courage, 2010).

A first version of the taxonomy was published in 2009 at the German GI Mensch & Computer Conference (Gerken & Reiterer, 2009c). In the following chapters, we will present a heavily revised and extended version of the taxonomy which provides a more detailed hierarchy of research questions and a more explicit interlinking between different parts of the taxonomy. When applicable, we provide example studies from the HCI context. For presentation reasons, we have subdivided the taxonomy into two major parts:

1. A taxonomy for research questions in longitudinal research in HCI
2. A taxonomy for research designs in longitudinal research in HCI

Appendix A provides an overview of the reviewed research papers in HCI and how they refer to our taxonomy.

2.2 A Taxonomy for Research Questions in Longitudinal Research in HCI

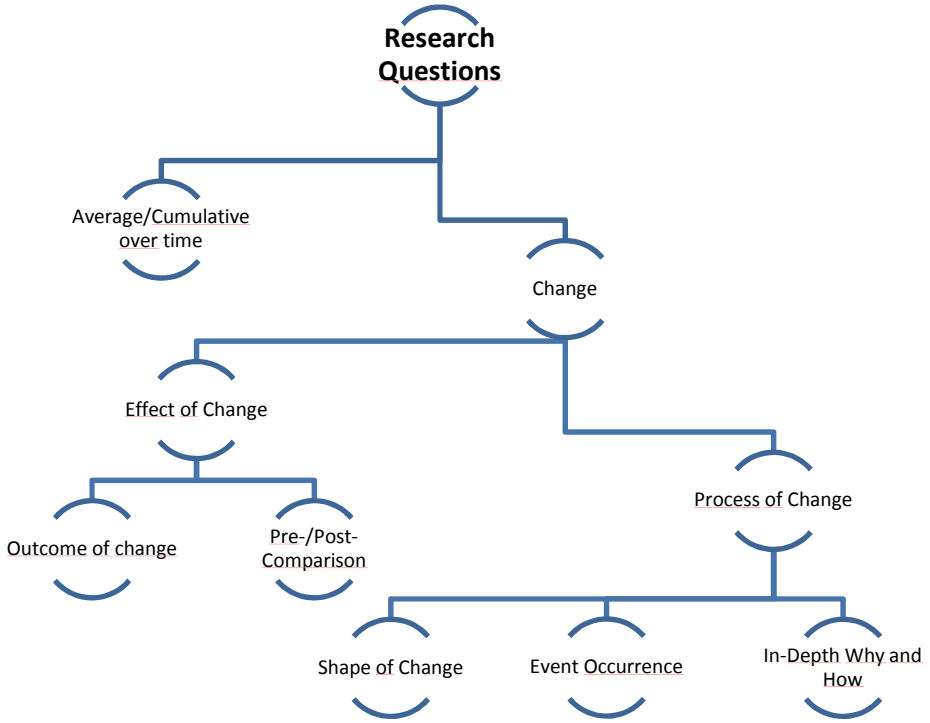


Figure 2: An overview of the taxonomy for research questions

Figure 2 depicts an overview of the taxonomy for research questions in longitudinal research in HCI. We will discuss in detail the different aspects in the following sections. The taxonomy encompasses two main branches: Interest in change and Interest in averages or cumulative data over time. Interest in change is often entitled as “true” longitudinal research while interest in averages or cumulative data seeks to answer cross-sectional research questions in a longitudinal setting. Accordingly, as our interest is mainly in “true” longitudinal research, we continue to branch the taxonomy for interest in change. We can then distinguish between Interest in the effect of change and Interest in the process of change. Eventually, the leaves provide the links to existing and appropriate qualitative and quantitative (statistical) methods for analysis. To some extent, one could argue that the taxonomy is upside down: As methods for analysis are still rare or simply unknown to the researcher, the choice of analysis method

often defines the type of research questions that can be asked. However, as history has shown, if important research questions exist, appropriate analysis methods are likely being developed. As we see this taxonomy as one that should be extended or refined over time, an organizational structure based on analysis methods would restrict our thinking and make it difficult to include questions for which no analysis method yet exists.

2.2.1 Interest in the Averages or Cumulative Data over Time

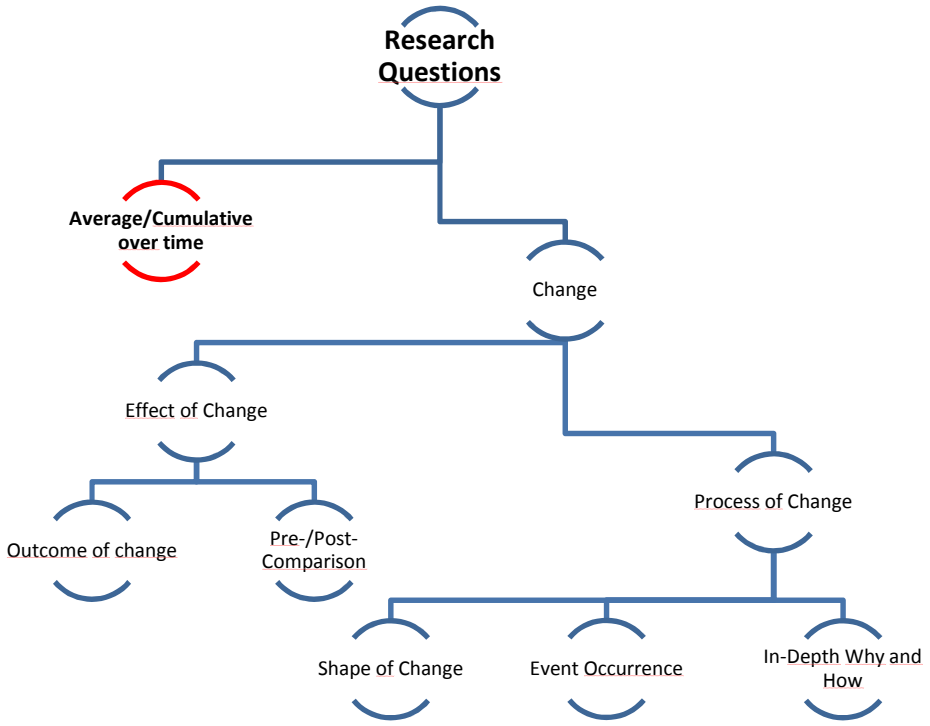


Figure 3: Interest in the averages or cumulative data over time

This first variant of research questions (see Figure 3) is not interested in change *per se* and is therefore not generally regarded as a longitudinal research question in other fields (e.g., (Menard, 2008), (Singer & Willett, 2003), (Bijleveld & van der Kamp, 1998)). Nevertheless, it is common practice in HCI to still call such studies *longitudinal*, as they share the characteristic of gathering data at multiple points in time. However, time is viewed merely as a factor that influences the reliability and thereby also the validity of the measurement – are we

really sure that what we see in the study is what happens in the real world? We call this the peephole dilemma of the cross-sectional paradigm. Just as looking through a peephole gives only a limited view of the room beyond, what we see as researchers when conducting a cross-sectional study is a very limited snippet of reality. We lack contextual information and are in danger of recording biased data because we are unable to capture the variability in the data over time. Furthermore, as time is often a limiting factor in such studies, we may end up with rather artificial settings. To illustrate this issue in detail, we will outline an example study before presenting real-world examples from HCI.

In a hypothetical study A, let us assume that we are measuring the usability of a novel interaction design through task time, error rate, and usability defects. We are doing so in a one-hour lab-based controlled experiment with 24 participants – a reasonable (or at least common) N for such experiments in HCI. In the end, our results include estimates for the average task time and error rate and the respective variances, as well as a number of usability defects. Interpreting these results involves several challenges. First, as time was limited, we were unable to include tasks for all features of the interaction design but instead had to choose a sensible subset. Furthermore, we could only include one measure of task time and error rate per participant. Finally, we do not really know much about the activities our participants were involved in prior to the study. The first aspect means that we are limited in what we can conclude about the complete interaction design. The second and third aspects increase the variability in the data. For example, having only one task-time measure per participant means that outliers can have a large effect on the results. It may be very difficult to interpret such results, as we have no additional data from these participants. Similarly, usability defects may be missed or over-represented if participants have worked with the design just for one hour. Users of real products often come up with innovative strategies or workarounds if a system does not behave exactly as anticipated – something that is unlikely to happen within one hour of observed testing with pre-defined tasks.

For study A, applying a longitudinal paradigm is not an easy resolution for these problems. However, it is the most straightforward approach to gathering more

data for each participant, thereby allowing us to increase the meaningfulness of our data and the reliability of our measurements. Bolger et al. (Bolger, Davis, & Rafaeli, 2003) describe these kinds of research questions as, “What is the typical person like, and how much do people differ from each other?” In our study A scenario, this would translate to, “What level of performance in terms of task time and error rate can the typical user achieve with the interaction design, and how much do people differ from each other?” At this point, from an analysis perspective, this is the same as a cross-sectional study. As a result, this longitudinal setting is frequently found in HCI research. From the 42 studies we analyzed in detail, 15 applied a longitudinal design mainly to capture more data and answer cross-sectional research questions in a more reliable way. As such, it is also a relatively established approach within HCI research. In particular, field studies (such as workplace studies) are often longitudinal in nature: people are observed over several days or interviewed several times. However, without proper framing of research questions and data-gathering, only cross-sectional research questions can be answered. Consequently, this also means that the data analysis is in principle comparable to cross-sectional studies.

2.2.1.1 Analyzing the Usability of an Activity-Based Computing Tool (Voida & Mynatt, 2009)

In their study, Voida & Mynatt were interested in seeing how an activity-based computing tool would be adopted by knowledge workers. Their tool, Giornata, was a prototype system that basically replaced the Mac-OS X desktop and allowed functionalities such as tagging, virtual desktop management, and activity management. The authors argued that a lab-based study in a controlled environment would not be able to capture the way knowledge workers would adopt such a system, as the setting would “limit the diversity of information, organization, and tools that participants could draw upon.” They therefore recruited five participants who were willing to carry out all of their computer-based work within the Giornata system. The ultimate goal of the study was to explore “how activity-based tools are adopted and utilized in real-world, authentic work environments and in the broader context of existing knowledge work artifacts.”

Participants used the system for a minimum duration of three weeks, after which the authors conducted a midpoint interview to assess usage styles. Participants were then allowed to leave the study; however, some continued up to a maximum of 82 days, undergoing a second interview after an additional 2 months of usage time. The average usage time was 54 days. In addition to these two interviews, the Giornata system logged typical interactions, such as activity creation, tagging, activity switching, etc.

Analysis and Results

The analysis was conducted in a typical manner for this type of research questions: averaging across time. The authors were thereby able to find out that on average, participants had 7.6 open activities and 28.2 switches between activities per day. As the latter figure varied to a great extent among participants, individual work styles based on the logging data were identified. One participant, for example, maintained a long and detailed list of activities, but switched only within a small subset of them. The interviews included subjective rating of the interface using 5-point rating scales. Overall, the system was liked (rated 4.2 on the 5-point scale) and collaboration support was identified as a possible bottleneck (rated 3.4 on average). In addition, the authors presented anecdotal evidence from the interviews, describing how participants used the system. The basis for the analysis here was a framework of challenges for activity-based computing. This framework had already served as the guideline for the design of the system and was now used to structure the interview data. Some anecdotes presented also address a kind of “change” research question, as participants commented on how the Giornata system behaved differently from their normal file system and how this changed the way they worked with their data. Still, there was no systematic analysis of how usage changed during the course of the study itself.

This study shows very well how applying a longitudinal research paradigm allows researchers to gain much more detailed knowledge, greatly increasing the validity of the data. It also shows that while it might have also been interesting to look for changes in usage patterns, this type of cumulative analysis can be appropriate for longitudinal research.

2.2.1.2 Analyzing the Usability of LiveRAC, a Visualization System for the Analysis of System Management Time-Series Data (McLachlan, Munzner, Koutsofios, & North, 2008)

McLachlan et al. present an interactive system called LiveRAC, which helps analyze time-series data by providing interactive visualization techniques. They address a very specific user group: system management professionals in the senior operator staff, so-called Life-Cycle Engineers. They describe their study as an "informal longitudinal study […] to better understand the strengths and weaknesses of the visualization techniques." The system was developed in a participatory design process, and the same 14 participants also took part in the evaluation. During the study, McLachlan et al. collected notes, audio, screen captures of desktop sharing, and log data. Interviews were done remotely over the telephone. They analyzed the data in a wiki to identify requested functionality and bugs, and observe user behavior. The findings are cumulative and describe anecdotal evidence illustrating how the different functions were used, which were preferred, and what problems occurred. Moreover, the authors illustrate three typical usage scenarios based on screen captures taken during the study. Similar to the study by Voida & Mynatt, the authors sometimes describe how the introduced system changed user behavior and what kinds of activities were not previously well-supported, but do not analyze change in a systematic way. However, the study helps to foster understanding of how the system was adopted in the real world; the design provided tremendous advantages over a short-term cross-sectional study design, as participants could work on real tasks with real data.

As we have stated in the beginning of this chapter, this kind of research question with interest in averages and cumulative data over time has little in common with longitudinal research that studies change processes. However, they do provide the researcher with a more balanced and reliable perspective on the data and as such, applying studies that incorporate these kinds of research questions are very common in HCI. However, one has to be aware that applying a longitudinal design without proper modeling of the time variable and without explicit analysis of possible change processes, might also lead to incorrect conclusions. In our study A, for example, the task-time performance might be

hugely affected by a systematic learning process. If such learning were present, we could not simply ignore it and average our data over time. The same is true for usability issues: we cannot assume that a usability issue encountered after the first 5 minutes is comparable in every aspect to one that only shows up after several hours of usage. Thus, doing this kind of longitudinal research increases the researcher's responsibility to sensibly interpret the data. We therefore recommend applying longitudinal research to questions with interest in averages and cumulative data over time only if no systematic changes are expected or if one takes them into consideration when interpreting the results.

2.2.2 Interest in Change

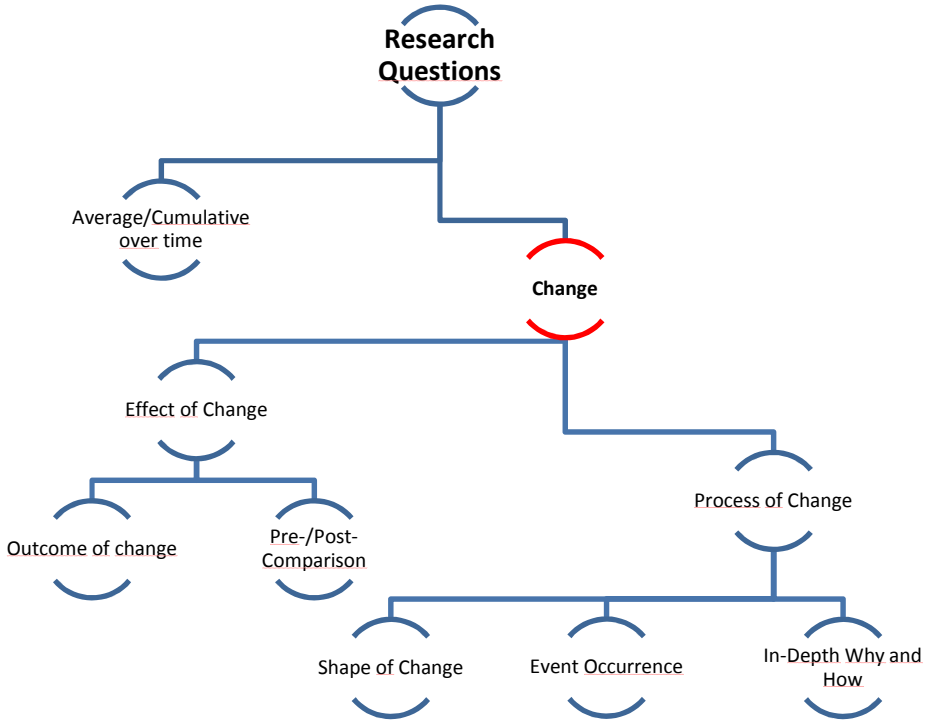


Figure 4: Interest in Change

While studies that apply a longitudinal paradigm to answer questions that show interest in averages or cumulative data over time are mostly useful for improving some of the shortcomings of cross-sectional research, interest in *change* opens up a whole new range of possible research questions. To our knowledge,

there is currently no classification scheme or taxonomy, even in the social sciences or psychology, that provides an overview of these research questions and how they differ. From an HCI perspective, one can clearly identify two main approaches. The first relates to research questions that show interest in the effect of change. This means that the change process itself is not studied, but instead measures are taken either to assess the outcome of a change process or to make before/after comparisons. On the other hand, there are also research questions that directly address the change process itself. From an analysis perspective, these are the most challenging questions, as we will later demonstrate in detail. Questions here encompass aspects such as the shape of change – linear (in the simplest case), or representing more complex behavior with ups and downs or even patterns. Event occurrences are also of interest, in terms of whether and when certain events occur. Singer & Willett (Singer & Willett, 2003) address these two issues from an analysis perspective and demonstrate that there are advanced statistical models available that allow statistical analysis beyond simple descriptive statistics: namely, multi-level growth modeling (for investigating the shape) and survival analysis (for event occurrence). While event occurrence captures qualitative data that can be analyzed in a quantitative way, the last aspect of this taxonomy focuses especially on qualitative research questions that try to go beyond statistical models and ask for the Why and How of in-depth change processes. For example, when we want to understand the process of technology adoption, the change processes cannot really be measured solely in numbers. Instead, such research questions ask the researcher to look for changes in observational or interview data, a difficult task. Here, we will give an overview of the analysis framework by Johnny Saldaña (Saldaña, 2003) (Saldaña, 2008), which helps and guides such qualitative studies. However, we think this is an area where more research is needed, not only to enable researchers to ask the proper questions but also to provide guidelines or frameworks for data gathering and analysis.

2.2.3 Interest in the Effect of Change

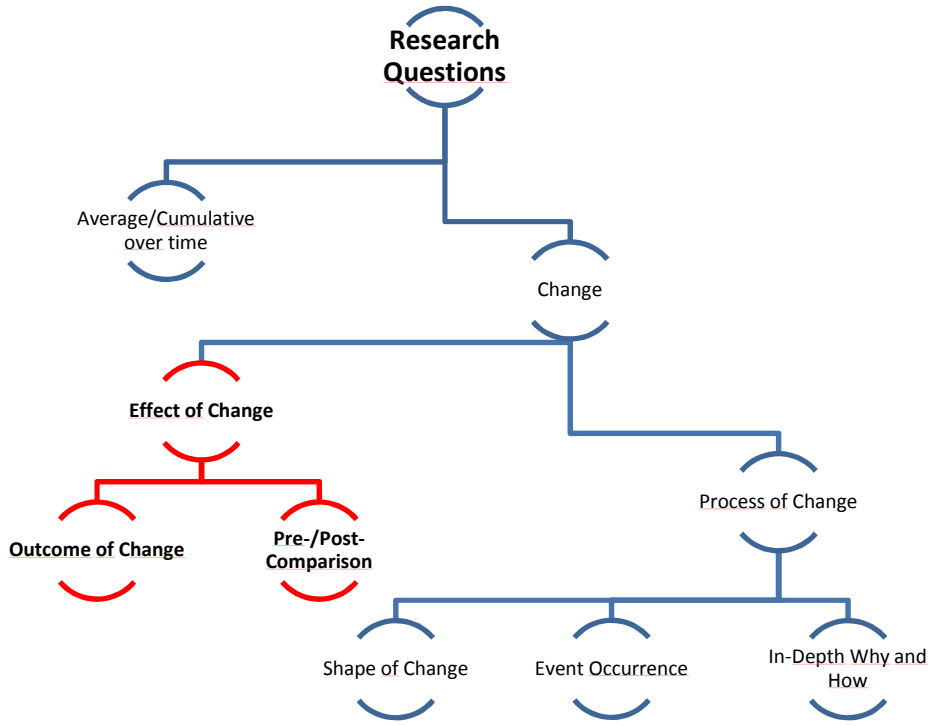


Figure 5: Interest in the effect of change

2.2.3.1 Outcome of Change

We will start by discussing the interest in the *outcome of change*. Such research anticipates change processes and might even monitor them, but the central research question is focused on the end product of this change process. One prominent example of this kind of research question is input device evaluation. Novel input devices often require an initial learning period, both for the motoric skill set to develop and for the user to understand how to use the device efficiently. As this learning process is very common, researchers are often not interested in the process *per se* (it is sometimes simply assumed that it follows the power law of practice – e.g. (Card, English, & Burr, 1978)), nor in a comparison with measurements that took place prior to the learning process. Instead, the researcher may be specifically interested in assessing the point in time when learning levels out, indicating the learnability of the device. Another related re-

search question might then ask how this particular input device compares to other devices.

Therefore, one important aspect of this research question is to actually define the point in time when the change process has finished or leveled off. For example, with quantitative data, Helmert contrast analysis has been used in several different studies (e.g. (Douglas, Kirkpatrick, & MacKenzie, 1999), (MacKenzie, Kauppinen, & Silfverberg, 2001), (Bieg, 2008)). Helmert contrasts compare the performance of each measurement session with the mean of all following sessions. Thus, if your longitudinal design included six measurement sessions, it would compare the first session with the next five sessions, the second session with the next four, and so on, providing a test of significance for each of these comparisons. Compared to simple pairwise comparisons, this procedure is better suited to taking into account the entire learning process. When using pairwise comparisons, outlier sessions (e.g., a participant having a bad day) have a stronger effect and make it more difficult to interpret the results. In pairwise comparison, one can easily encounter a situation in which learning stops from one session to the next and then "starts" again. Helmert contrasts on the other hand are much more conservative and react slower to such fluctuations in the data. Thereby, it may represent a lower bound for the outcome of a change process.

Some example research questions for interest in the *outcome of change*:

- Do participants using novel input device A (laser-pointer) achieve better pointing performance (e.g., Fitts' Index of Performance) compared to the established device B (mouse)?
- After having bought an iPhone, do users immediately (within the next 2 days) start buying apps in the market?
- After having purchased an iPhone and used it for at least 3 months, how do people approach touch-based devices in public?

2.2.3.2 Pre-Post Comparisons

Interest in *pre-post comparisons* is closely related to the research question about the *outcome of change*, and often both research questions are addressed within one single study. This research question asks for a before/after comparison and an assessment of what and how much has changed. In contrast to the outcome of change question, it is not necessary to make sure that the change process is completed and that measurements are stable. Instead, there might be good reason to compare before/after simply by a fixed amount of time. A before/after comparison therefore does not preclude any additional changes, but simply assesses the amount of change that has taken place over a specific time period. In principle, most experimental designs also incorporate this kind of research question. Given an experimental treatment (that happens over time), changes are observed and analyzed. The longitudinal approach, however, allows the inclusion of longer time periods and measurements or treatments in between the before/after framing of the study. From an analysis perspective, interest in the size of change is quite easy to analyze in the case of quantitative data. In general, widely used methods such as repeated-measures ANOVA or pairwise t-tests can be applied. In the case of qualitative data, pre-post comparisons are naturally more problematic, as it is more difficult to assess changes. One common possibility is quantification; in the simplest case to coding event existence with 1 or 0, such as whether or not usability problems do occur before and after.

Example research questions for interest in the size of change:

- Do people rate their mobile phone's attractiveness and usability significantly differently after having used it for more than 2 months?
- Do people perform significantly better with a laser-pointer (Fitts' Index of Performance) after having used it 30 minutes per day for one week?
- How much can the error rate be reduced by providing two weeks training on a new accounting system?
- Do people face the same usability problems after having worked with a new accounting system for 6 months?

In the following two sections, we will describe two example studies – one from our own experience and the other from the literature – to illustrate research questions with interest in the *outcome of change* and in *pre-/post-comparisons*.

Laser-Pointer Performance over Time (Bieg, 2008) (Gerken, Bieg, Dierdorf, & Reiterer, 2009a)

As a more detailed example, we will illustrate the combination of interest in the outcome of change and interest in the size of change in a longitudinal experiment that investigated learning to use a laser-pointer as an input device for large high-resolution screens and how performance compared to mouse input. It was a small-scale study, so the focus here was not on the validity of the results, but rather on illustrating how to address such research questions.

Figure 6: Study Setup: Multi-directional tapping task (green bubble represents target object)

In this study, we applied a longitudinal panel design with five data-gathering waves on five consecutive days. The experiment took place in a lab, as neither laser-pointer input devices nor large high-resolution displays are commonly used. We selected six subjects to use the laser-pointer on five consecutive days

for 30-45 minutes each day. The practice task followed a discrete multi-directional tapping paradigm (see Figure 6 but was enhanced with a feedback component to keep users' motivation high (similar to the study by Card et al. (Card, English, & Burr, 1978)). Each session consisted of 756 trials per participant. In the first and last sessions, participants completed an additional different experimental task, a continuous one-directional tapping task designed to distinguish task learning from input-device learning. Two blocks were performed with the laser pointer (marked as OL in Figure 7) and two additional blocks were performed using a mouse (OM). We assumed that the performance between the first and the last sessions 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. We formulated the following research questions:

- Interest in outcome:
  - Q1: How long does it take participants to learn to use the laser-pointer device?
  - Q2: How does a laser-pointer compare to mouse input, in the case that participants are provided with practice sessions with the laser-pointer (as they are not familiar with its use)?
- Interest in pre-post comparison:
  - Q3: Does the performance in terms of index of performance and movement time significantly increase over time when participants are provided with practice sessions in between measurements?
  - Q4: If yes, how large is this performance increase?

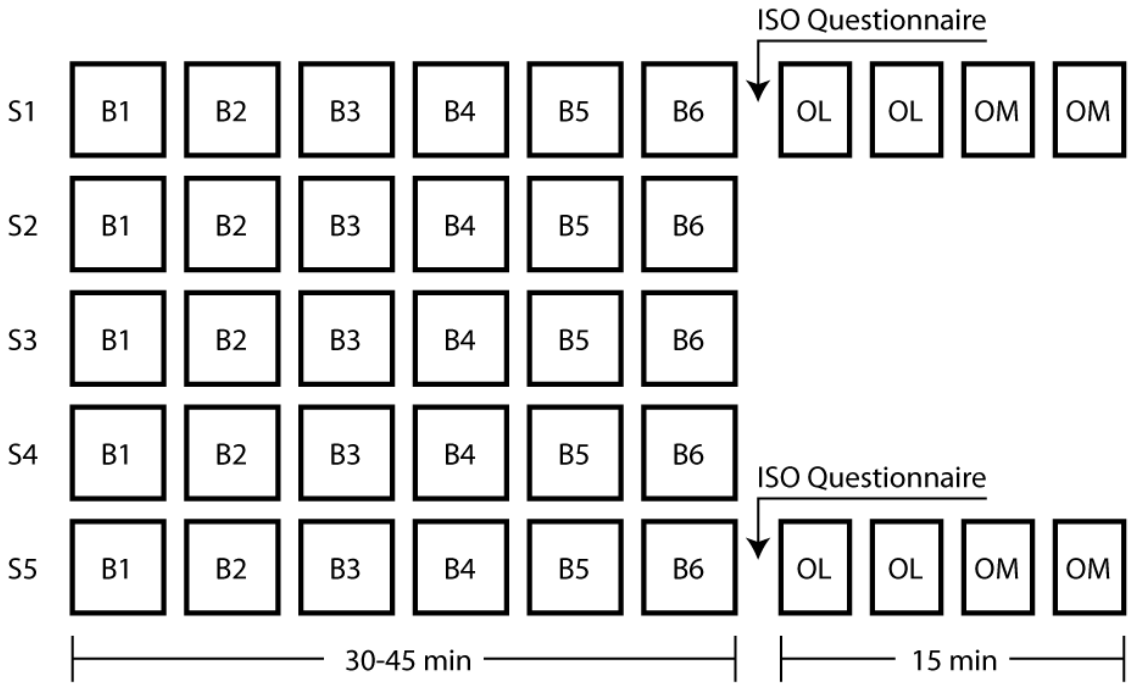


Figure 7: Data-gathering design for the longitudinal laser-pointer study

Analysis and Results

Based on the benefits discussed above, we used post-hoc Helmert contrast analysis to assess how long it took participants to learn to use the laser-pointer device. This procedure compares the performance of one session with the mean of all following sessions. The analysis shows that performance improved significantly up to the fourth session (see Table 1). Afterwards, performance dropped slightly, although this drop was not significant. Thus, we can conclude that the learning process took our participants approximately 4 sessions of 30 minutes, or 3024 trials.

As each session consisted of 756 trials, this analysis gives a rather rough estimation. In addition, we cannot rule out the possibility of Session 5 being somehow an outlier session or that additional practice could lead to even better performance.

Table 1: Helmert contrast analysis

	F(1,5)	p-value	Partial Eta Squared
Session 1 vs. later	290.19	<0.001	0.98
Session 2 vs. later	16.27	0.010	0.77
Session 3 vs. later	10.15	0.024	0.67
Session 4 vs. Session 5	0.11	0.752	0.02

To address the second research question, we compared the performance with the laser-pointer with the mouse after Session 5. As discussed above, we separated the practice task from the experimental task for comparison, thereby reducing a possible task-learning effect. As a consequence, we cannot draw a comparison between the laser-pointer and the mouse for the point in time at which our Helmert contrast analysis shows that learning leveled off (Session 4). Instead, we can make this comparison only for the first and last sessions, with the latter being of interest here. Results show that the laser-pointer performance was significantly worse than the mouse input (4.18 (sd: 0.42) vs. 4.61 (sd: 0.32) bits/s, $t(5)=-2.81$, $p=0.037$, 5% level of significance). With regard to the size of change in performance due to learning, it may be of additional interest to compare this difference with the difference at the beginning of the experiment. The difference between laser-pointer and mouse was on average 0.43 bits/s in Session 5 (sd: 0.37), in comparison to 0.71 bits/s in Session 1 (sd: 0.32). A paired-sample t-test shows that this difference is significantly smaller in Session 5

compared to Session 1, demonstrating that learning significantly improved performance with respect to a comparison input device (t(5)=3.64, p=0.015).

Analyzing the overall learning effect revealed that participants improved their performance significantly (3.83 to 4.18 bits/s, SD: 0.38 vs. 0.42, t(5)=-4.132 p=0.009), while the mouse performance remained stable (4.54 compared with 4.61 bits/s, SD: 0.29 vs. 0.32, t(5)=-1.23, p=0.272). Figure 8 illustrates these differences graphically.

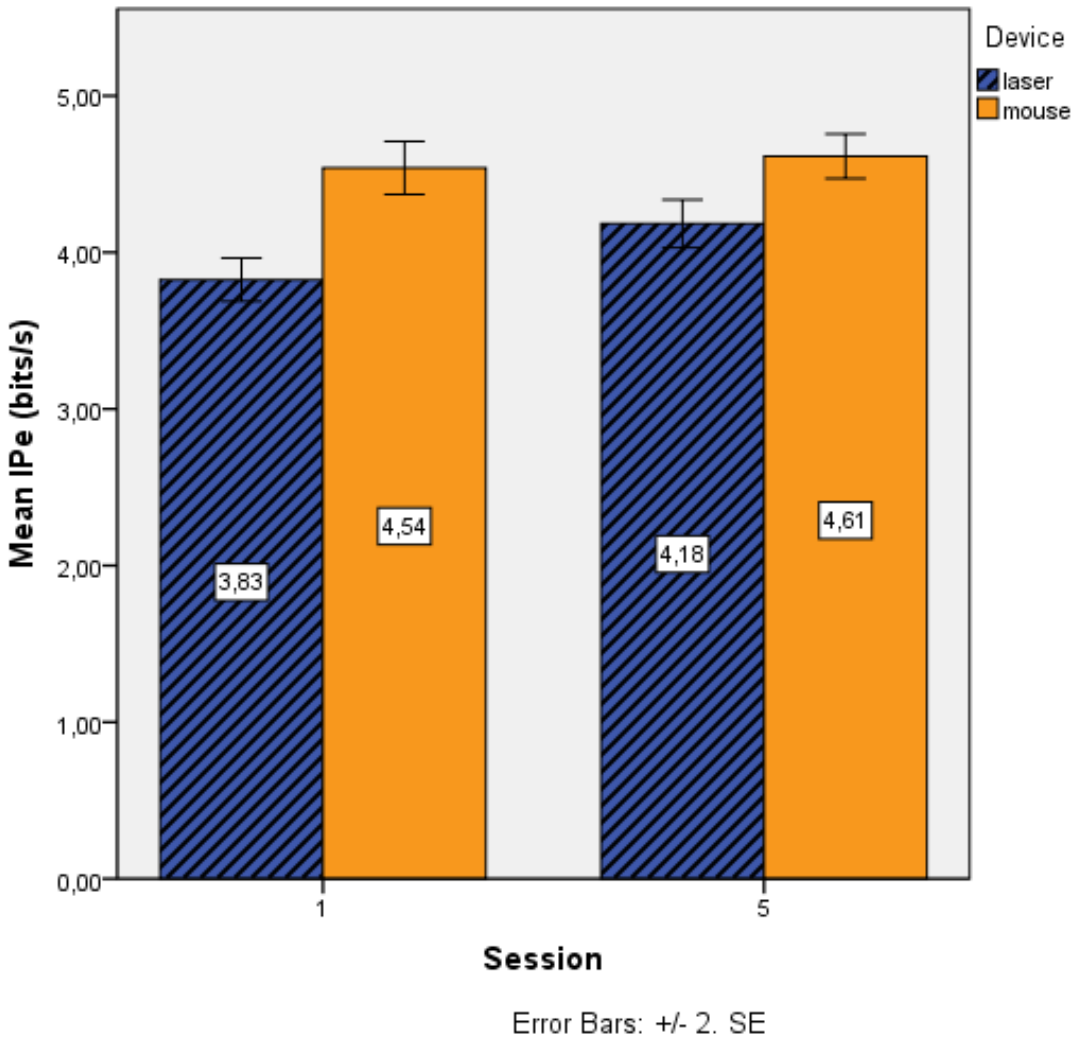


Figure 8: Analyzing the size and outcome of change for laser-pointer performance

With respect to our research questions, we were able to answer them as follows:

- Q1: It took our participants on average four sessions (of 30 minutes and 756 trials each) to complete the learning process. As discussed above, this result should be regarded with caution, as data beyond Session 4 is only available for one additional session.
- Q2: The laser-pointer performance was significantly worse compared to the mouse, both prior to and after practice. However, the magnitude of this difference is significantly smaller after practice.
- Q3: There is a significant increase in pointing performance over time, from an average of 3.83 bits/s to 4.18 bits/s in the experimental task.
- Q4: The average increase is about 9.1% or 0.35 bits/s. It was higher during the practice task itself (nearly 2 bits/s) but as expected this includes the task-learning effect, which we were able to isolate by selecting different experimental tasks.

Does Time Heal? Usability Problems in Pre-Post Comparisons (Kjeldskov, Skov, & Stage, 2005)

The second example study was conducted by Kjeldskov et al.. It clearly illustrates the pre-post comparison, focusing on changes in usability problems over time. The researchers analyzed how users of an electronic patient record system in a hospital went from being novices to becoming experts. However, the change process itself was neither monitored nor analyzed; instead, the study focused on two points in time – when the system was introduced into the hospital and all participants were novices, and 15 months later, when the same group of participants had acquired a significant level of experience and could be regarded as expert users. The research question we will address here was:

- Q1: “Which usability problems are experienced by novices and by experts: which problems are the same and is there a difference in the severity of the problems that are experienced by both novices and experts?” (Kjeldskov, Skov, & Stage, 2005)

In addition, the authors investigated workload and usability measures such as effectiveness and efficiency.

Seven nurses participated in the study. While the exposure of the participants over time happened in the field, the measurement sessions at the beginning and after 15 months took place in the lab. After 15 months of usage, the participants indicated that they had used the system about 2 hours per day and were consequently characterized as expert users. The study implemented a classic usability testing paradigm with think-aloud protocol and users completing typical tasks on the system for about 45 minutes. The same tasks were used in the two measurement sessions. Measures included task completion time, workload through means of the NASA TLX (Hart & Staveland, 1988), and the identification of usability issues.

Analysis and Results

Two important aspects of the data analysis process here were that 1) data was analyzed by two researchers who did not act as test monitors during the lab sessions and 2) the researchers analyzed the video recordings of the sessions randomly, without knowing which of the two measurement sessions the recording belonged to. In this way, they tried to avoid any subjective bias on the part of the analyst when assessing the type and severity of the usability issues. The severity ratings (cosmetic, serious, critical) were based on an individual level instead of a global rating for each usability issue. This allowed the authors to better analyze whether usability issues were perceived differently or had a different severity impact when encountered by expert users in comparison to novice users.

The results showed that many of the problems endured over time. Overall, 43 of 103 usability problems were encountered both at the beginning and after 15 months of usage. In addition, a number of the 103 usability issues were unique problems to one individual participant. Omitting these, 40 out of 61 problems persisted. Interestingly, most of the critical problems remained as well (17 out of 21). However, the individual severity ratings showed that overall the shared problems were seen as less severe. The authors used a Wilcoxon signed rank test to examine this difference for significance. This analysis method is a non-parametric counterpart to the paired t-test and is well-suited to analyzing repeated measurements (as in longitudinal designs) when a normal distribution

cannot be assumed. The mean value for the novice severity score was 1.91 (SD=0.51), and for the experts 1.55 (SD=0.57). The test results showed that this difference is significant (z=3.963, p=0.001). The authors conclude that “a remarkably high number of problems were experienced both by novices and expert users. These problems were experienced significantly more severely for the novices, so the problems that remained became less severe” (Kjeldskov, Skov, & Stage, 2005).

2.2.4 Interest in the Process of Change

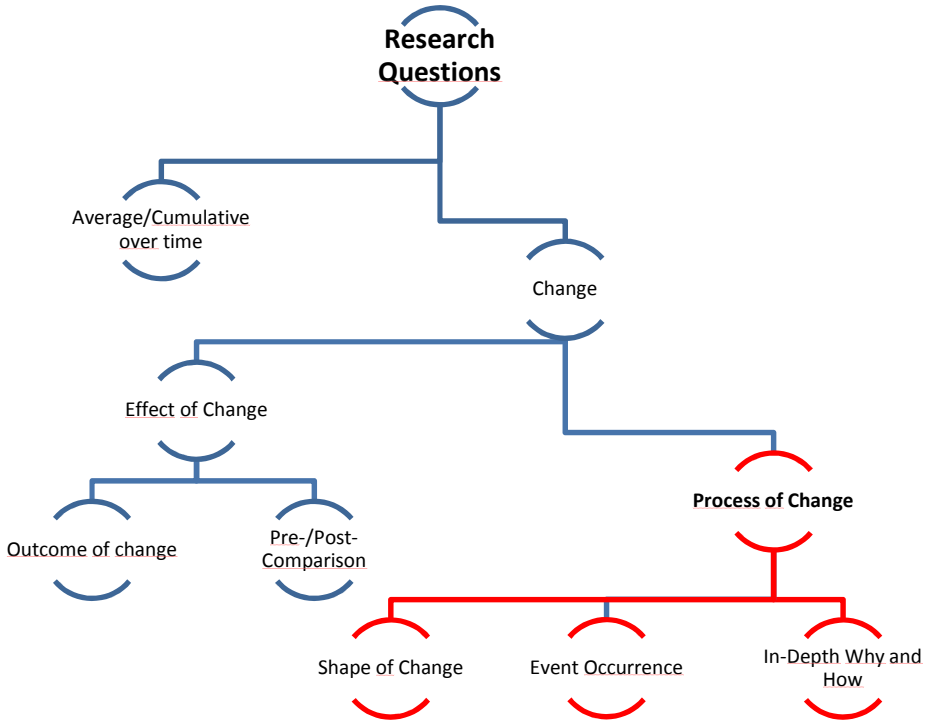


Figure 9: Interest in the process of change

All of the research questions to this point have shared many similarities with cross-sectional research. For example, pre-post comparisons are quite similar to controlled experiments featuring a within-subjects factor with at least two levels, both substantially and from an analysis point of view. Research questions that analyze *the process of change* are different, in that they do not simply analyze results of change processes. Rather, they disassemble the process itself to

analyze the shape of the change, to identify whether and when certain events occur, and to understand the change process in detail. Obviously, analyzing longitudinal data in this way is much more difficult, which probably explains why many studies in HCI feature mostly descriptive statistics, graphical representations, and anecdotal evidence that help the reader to get an understanding of the data. Analytic frameworks that facilitate understanding the processes from a qualitative point of view are also still rare, making the analysis of changes in qualitative data an even more challenging task.

2.2.4.1 Interest in the Shape of Change

One way to understand the process of change is to look at *the shape of change*. By shape, we mean the overall pattern reflected in the measure of change. For example, we could analyze how the user experience rating assessed by a questionnaire such as AttrakDiff (Hassenzahl, Burmester, & Koller, 2003) changes over time: whether it increases or decreases continually or has ups and downs. Analyzing and understanding the shape can help us to discover certain time-dependent patterns in the data. It may also help to identify potentially interesting points in time, as large changes could indicate a certain important event that might help to explain the pattern. Furthermore, understanding the shape of change adds explanatory power to pre-post comparisons, as it sheds some light on the time in between the pre and post measurements. Eventually, understanding the shape of change might even allow us (to some very limited extent) to predict future changes. However, analyzing the shape of change can also lead to faulty conclusions. Measurement errors or natural variability might simulate nonexistent changes. Therefore, one must be careful when interpreting even the slightest bending in the shape of a change process. Examples of research questions that address the shape of change include:

- How did the usability rating of the iPhone change over time? Did it continually rise or decline? Was it stable? Were there ups and downs?
- Is it possible to learn input device A faster than input device B?
- Does the usability rating of an iPhone change differently from that of a competing Android phone over time?

- How are newly discovered usability problems of a website distributed over time? Are most of them found in the beginning? Are they evenly distributed over time?

In order to analyze the shape of change, we need at least three measurements over time and if possible many more, to reduce the danger of over-intellectualizing the data (Singer & Willett, 2003). In Human-Computer Interaction, statistical analysis of the shape of change is rare. Most studies focus more on an exploratory analysis, either by presenting graphical representations of the shape of change or anecdotal evidence for change patterns in qualitative data. One reason for this might be that the necessary statistical methods not "common knowledge". Even in other disciplines, researchers have struggled for quite some time to find appropriate statistical models and approaches. However, nowadays a variety of approaches have been adapted to fit longitudinal data, such as Structural Equation Models (Bijleveld & van der Kamp, 1998, p. 207ff), Latent Class Analysis (Dayton, 2008), Generalized Estimating Equations (Hilbe & Hardin, 2008), or Logistic Regression (Menard, Panel analysis with logistic regression, 2008). One approach we found especially interesting is the multilevel growth modeling approach (Luke, 2008), which is very intuitive in the way it treats longitudinal data but is nevertheless extremely powerful and flexible. In several leading books about longitudinal data analysis (Singer & Willett, 2003) (Menard, Longitudinal Research, 2002), this method is regarded as one of the best if not the best approach to obtain insight into this type of research question. After outlining two example studies from HCI, we will briefly introduce the basic assumptions and procedures of such a method. The interested reader is referred to the excellent book by Willet and Singer.

Laser-Pointer Performance over Time (Gerken, Bieg, Dierdorf, & Reiterer, 2009a) (Bieg, 2008)

As a first example, we will refer back to the laser-pointer study described for the research question of the outcome of change and pre-post comparisons. We are thus able to illustrate that it is not uncommon for a study to simultaneously tackle multiple types of research questions from the taxonomy. For our current interest in the shape of change, this study illustrates how plotting data over time

can help to analyze and understand the shape of change. To review, participants in this study used a laser-pointer for a pointing task during five sessions on five consecutive days. An experimental task before and after these five sessions was used to assess pre-post differences and to allow comparison to a common input device. In Figure 10, we can see the shape of the learning process over the five days. During each session, six data-points were taken (blocks); the vertical bars mark the transition from one session/day to the next. In (Bieg, 2008) this graph is analyzed in depth. What we can see here is that there was an overall performance increase for all participants over the five days. Interestingly, for some users (ID1-4) this increase happened right at the beginning, during the first 3 or 4 blocks. From the end of Session 2 on, performance seems to have reached a peak for these users, while others continually increased their performance up to Session 4 (e.g., ID4). Another interesting aspect is the drop in performance at the beginning of nearly every session, indicating that participants had to get used to the device again. In addition, it can be observed that most users did not achieve their best performance during the last block of each session, but instead peaked more towards the middle, which could indicate fatigue.

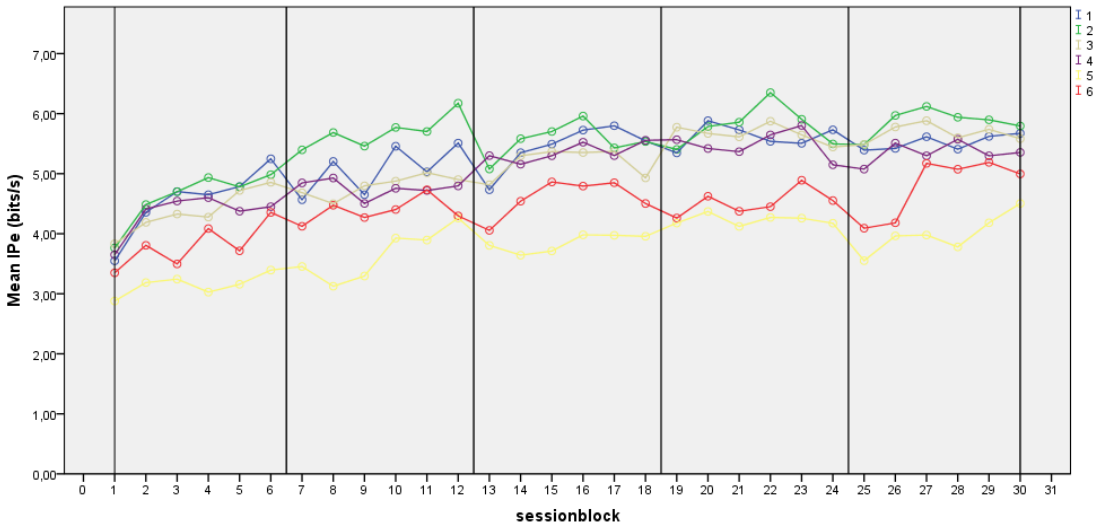


Figure 10: Performance development over time for six individual participants

The shape of this change process already tells us a great deal about how the device was learned and how to interpret the data. We can also see that there is

quite a lot of variability within participants within sessions, also indicating that we should gather data from more participants to reaffirm our results.

Usability over Time (Mendoza & Novick, 2005)

One of the most comprehensive papers on assessing the shape of change with qualitative data is the study by Mendoza & Novick on usability over time. These researchers had the opportunity to conduct a longitudinal study that focused on the usability of a "Home Page Designer" application, which was provided to the faculty at a middle school. The researchers were especially interested in the following research questions (citations from (Mendoza & Novick, 2005)):

- "Do users' levels of frustration caused by usability problems change as a function of experience with an application?"
- "Do the kinds of usability problems users encounter with a new system change over time as a function of use?"
- "Does the way that users respond to usability problems change over time?"

During an eight-week period, 32 teachers worked on an assigned test project that asked them to create a website to communicate information to students and their parents, such as student projects, homework, student work, and general class information. The task was sub-divided into smaller sub-tasks that built upon each other; participants were given these sub-tasks once a week during training sessions. The authors of the paper created the tasks and supervised the training sessions. Thus, while it was conducted in the field, the study still retained a relatively high level of control. For data-gathering, the authors prepared a "post-frustration experience survey," which was filled out by the participants whenever they became frustrated during use of the system. Such a design resembles the diary method approach, which we will discuss extensively in Chapter 3. The questionnaire asked participants to rate their current level of frustration and their self-assessed proficiency with the software on a five-point rating scale. In addition, they were free to supply reasons for the frustration and to indicate whether and how they solved the problem by marking a solution from a pre-defined set of choices (e.g., "asked someone for help").

Analysis and Results

Mendoza & Novick analyzed their data in a variety of ways. In a first step, they classified all 243 frustration reports they received from all 32 participants over the eight weeks. They adapted a classification scheme from the existing literature in this area and coded the frustration reports independent of their emergence in time. To ensure reliability, multiple researchers conducted this coding procedure independently. In the next step, they plotted this data over time. Figure 11 shows the different episodes and when they occurred during the eight weeks. With this data, the researchers were already able to draw important conclusions. For example, the high number of “user errors” in the beginning could be problematic in cross-sectional usability tests, as these problems do not seem to play a major role later on. Additionally, the high peak of “hard to find features” in Weeks 3 and 4 with a huge drop immediately following is interesting, as it might indicate that users needed some time to encounter the more difficult tasks and then got into trouble, but at some point had developed a base set of known functions.

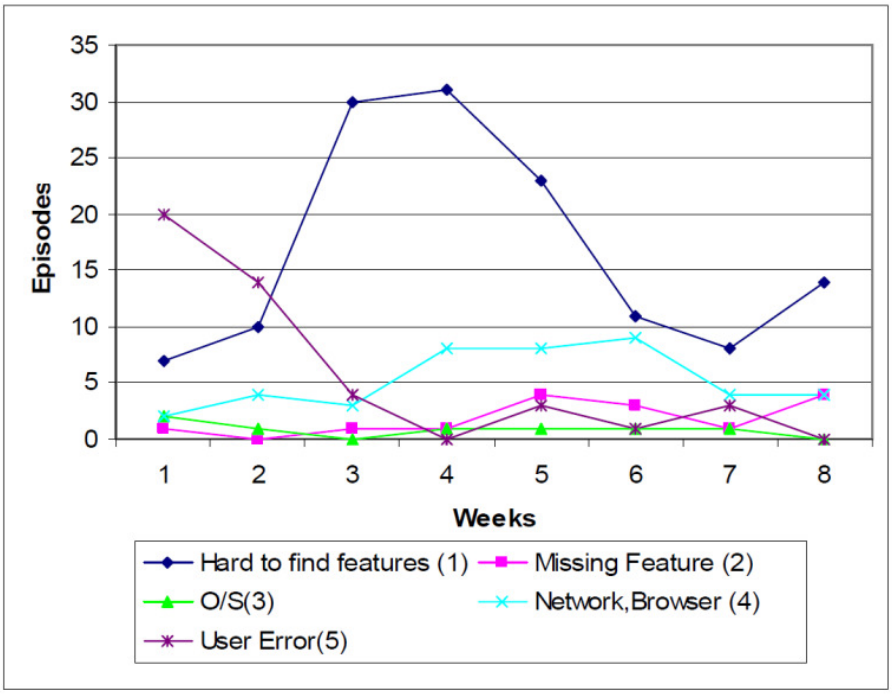


Figure 11: Frustration episodes over time (taken from (Mendoza & Novick, 2005))

Another aspect of the analysis process was the self-assessment ratings of proficiency and frustration level. Table 2 shows the weekly averages for these two measures. Even without a graphical representation, it is easy to see that over time proficiency goes up while frustration goes down. The authors used a repeated-measures test (a paired t-test, we assume) to analyze the data and found that frustration decreased significantly over time. With the help of multilevel growth modeling techniques, it could have been interesting to see whether this drop-off in frustration was linear and also whether proficiency could be a potential predictor of any differences in the individual shapes among participants.

Table 2: Proficiency and frustrations level averages over time (taken from (Mendoza & Novick, 2005)

Week	Proficiency Averages per Week	Frustrations Level Averages
1	1.4	3.8
2	1.2	3.9
3	1.6	3.6
4	1.9	2.6
5	2.1	2.2
6	2.3	1.6
7	2.1	0.9
8	3.4	0.8
Overall averages	2	2.425

Figure 12 adds yet another perspective to the data: it analyzes the responses users gave for their frustration episodes and how these responses changed over time. The authors plot the relative number of incidences, not the absolute numbers. This is a sensible choice here, as otherwise it would be not possible to compare the different responses over time (as the number of frustration episodes trends downward over time).

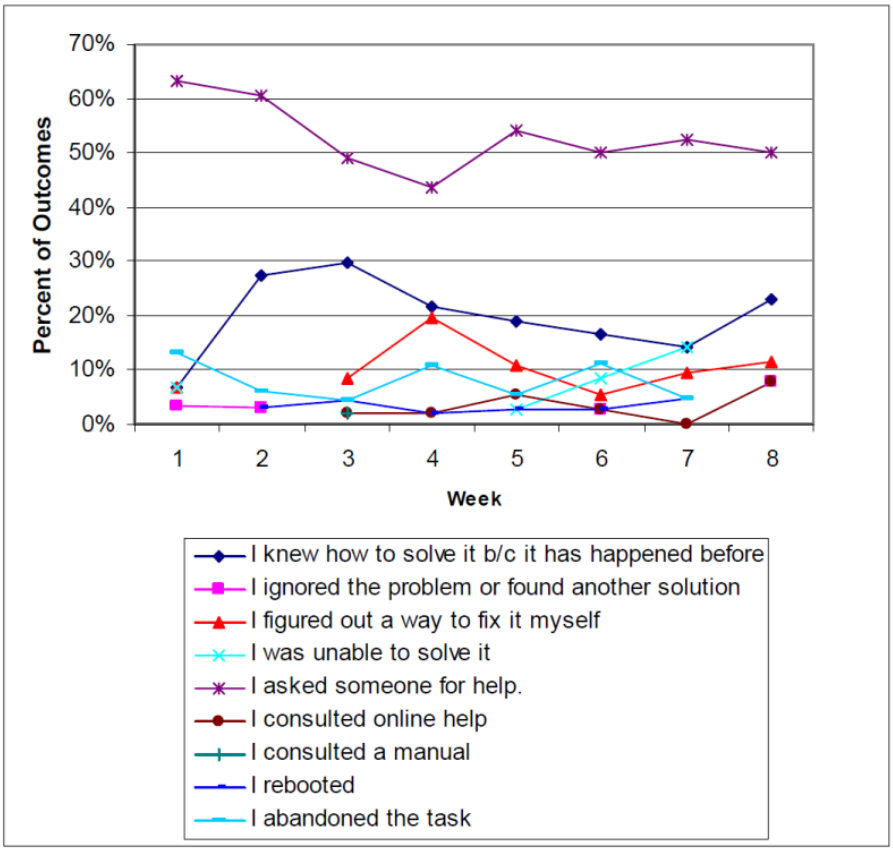


Figure 12: Relative Incidences of Users' Responses to Frustration Episodes (taken from (Mendoza & Novick, 2005))

An interesting pattern here is the fact that people only started to figure out ways to fix problems by themselves in Week 3, but not before. Also interesting is the fact that "asking someone for help" was the predominant solution to frustration episodes throughout the entire time period, although one might expect this to be more prominent in the beginning and to decline over the course of the 8 weeks as participants got to know the system better. Again, it would be interesting to see how this panned out on an individual level and whether some of these aspects might predict different frustration level change shapes among participants. In all, this paper presents one of the most thorough analyses in HCI of change processes and the shape of change in particular.

Multi-Level Growth Curve Modeling – A Brief Introduction

We are unaware of any longitudinal study in HCI that applies multi-level growth curve modeling techniques, although they provide one of the most flexible and

powerful approaches to dealing with quantitative longitudinal data. The primary purpose of the method is to help the researcher to describe the form and structure of changes in a quantitative dependent variable over time (although ordinal data can also be used) and thereby to explore inter-individual (or time-independent) and intra-individual predictors for change (Luke, 2008). In effect, this allows us to explain how measures might change over time within a group as well as why this group might be different from another group. For example, in the experiment reported in (Gerken, Bieg, Dierdorf, & Reiterer, 2009a), this approach would have allowed us to include a mouse-control group and quantitatively compare the learning shapes (as a whole) between the two groups.

A basic assumption of the technique is that observations are nested within individual cases. This has been especially beneficial in educational research, where such nesting can be used to represent hierarchies of organizational structures: e.g., students are nested within classes, and classes are nested within schools (Bijleveld & van der Kamp, 1998, p. 269ff). The multi-level nature of the model allows differentiation of these different hierarchical levels during analysis. In addition, these models assume a dependency among the observations nested within each other – something traditional “single-level” models are not capable of. Instead, traditional models assume that observations happen independently – in the example given, however, the observations on a class level are not independent of the observations on the student level, as students make up the class. This fundamental difference makes of the multi-level growth curve models especially well-suited for longitudinal data in which data from one individual cannot be regarded as an independent measure. When applying multi-level growth curve modeling to longitudinal data, repeated measures over time are treated as observations nested within an individual, thereby forming a simple hierarchical data structure with individuals at the top and time-points at the bottom level (Bijleveld & van der Kamp, 1998).

In experimental research, this basically allows us to include within-subjects (for intra-individual change) as well as between-subjects factors (for inter-individual change) in the same design. An example of an inter-individual predictor in an HCI study could be the use of a control group using a baseline interaction de-

sign. An intra-individual predictor in a similar study would mean that the same group of participants would switch to a second interaction design after a predefined time span, similar to a standard repeated-measures experiment in a cross-sectional design.

In addition to the inherent assumption of data interdependency, multi-level growth curve modeling has a number of further advantages for longitudinal data analysis with focus on the shape of change. First, it can easily cope with missing data (e.g., participants who have not taken part in every data-gathering wave). Second, the data-gathering waves do not have to be evenly distributed among participants; instead, each individual can have a unique data-gathering schedule (as long as the time-variable is coded as real time). For field studies, it is often impossible to maintain a regular data-gathering schedule; however, data-analysis methods such as repeated-measures ANOVA require all participants to have the same schedule. An important difference between data-structuring and cross-sectional analysis methods is the need for a “person-period” (or “long”) data set (Singer & Willett, 2003, p. 18), which uses one row for each time-point of data-gathering and therefore several rows per participant, instead of the classical “person-level” (or “wide”) data set approach that shows all measures for a participant in one row. As a result, we can include an explicit time variable. This is important, as it allows inclusion of missing data (without having to exclude a participant who only missed one data-gathering wave) and also of varying data-gathering waves. It additionally allows much easier integration of time-varying predictors, which basically are all intra-individual change predictors.

A typical longitudinal study to be analyzed with growth curve modeling should include up to five different variables (and one column for each in the long format): 1) The ID that identifies the participants; 2) one or more longitudinal dependent variable; 3) one or more variables containing time information, in “real time,” such as age or usage hours, or “study time,” such as the data-gathering wave number; 4) one or more time-varying predictors to predict intra-individual change; and 5) one or more time-invariable predictors to predict inter-individual change.

- The Basic Model

  The model resembles a multi-level hierarchical regression model (Bijleveld & van der Kamp, 1998, p. 271). To enable analysis of the data for inter-individual change processes as well as intra-individual change processes, the model has two levels. The level-1 sub-model describes how individuals change over time (the intra-individual part) and the level-2 sub-model describes how these changes vary across individuals (the inter-individual part) (Singer & Willett, 2003, p. 47). The level-1 model creates a regression model for each individual participant, modeling the relationship between time and the dependent variable. The simplest case would be to assume a linear growth model; many authors (e.g. (Singer & Willett, 2003) (Luke, 2008)) argue that one should be careful about choosing much more complex models, especially if there are only few waves of data. Luke, for example, suggests not going further than a polynomial quartic term, unless there is a good theoretical foundation to do so. One such example might be the power law of practice, to which HCI researchers often refer when evaluating input devices (see (Bieg, 2008)).

  A basic level-1 model could look like this (from Luke 2008, p548):

$$Y_{ti} = \beta_{0i} + \beta_{1i}T_{ti} + \varepsilon_{ti}$$

  Here, the dependent variable Y is measured at time point t for an individual i. The only predictor in this example to account for change is time T. β_0 is the intercept and β_1 is the slope of the linear regression model. As all betas have *i* subscripts, this tells us that they can vary for each participant individually. The same is true for the error term, which can also vary for each measurement time-point. The error term accounts for the amount of change not explained by the linear model. By this separation for each individual, we are able to analyze intra-individual change processes. Further time-varying predictors (in addition to time) can be included here, as well as the interaction between different predictors.

The parameters of this level-1 part are the outcomes of the level-2 part:

$$\beta_{0i} = \gamma_{00} + \upsilon_{0i}$$

$$\beta_{1i} = \gamma_{10} + \upsilon_{1i}$$

Each parameter $\beta_{0i}\ or\ \beta_{1i}$ (intercept or slope) is predicted by the grand mean of all the individual intercepts/slopes, with the addition of the variability of the individual parameters around the grand mean $\upsilon_{0i}\ or\ \upsilon_{1i}$. Again, we can add predictors, in this case for inter-individual change.

- Analyzing the Data

  An important strategy for all longitudinal data analysis interested in the shape of change is to first explore the data by plotting it in graphs. In order to be able to explore inter-individual changes, one should do so at the participant level, meaning one plot over time for each participant (and each measure). Thereby, one can assess whether people change similarly or completely differently in terms of the shape and the amount of change. Furthermore, this is a requirement for fitting any kind of model to the data, e.g., in the above simple example, a linear change model. Singer & Willett also frequently point out that one should be careful not to over-interpret every single up-and-down shift in the data, as this is more likely to be due to measurement error than to “real” change. Another important aspect to consider is the centering of the data. This means that the time variable should basically start with 0 so that the model does not “predict” anything before the first measurement, and that the intercept resembles this first measurement. In the case of measurement waves as time values, this can be done on the coding level. In the case of real time values, such as age, this must be calculated within the model to prevent later misinterpretation.

  The basic linear growth model gives us two parameters that can then be compared within and between users: the intercept and the slope. The intercept, as noted, tells us what happened in the first data-gathering wave; the slope tells us how fast and to what extent things are changing. The tech-

nique then allows us to assess whether this differs over time within-subjects, based on an intra-individual predictor, or between different groups of participants, based on an inter-individual predictor. For example, we are able to tell whether the satisfaction with a given product decreases faster after a new model has been released (within-subjects) or faster than the satisfaction of participants using another product (between-subjects).

While a basic linear model should not be rejected lightly, more complex models are sometimes necessary. There are several techniques to test the model fit, although none of them is without flaws; examining the data in graphical form remains important (see (Singer & Willett, 2003) for more details here).

Overall, while the approach is still not a standard part of every statistics software package, it is both simple and powerful and would allow for more complex analysis in longitudinal studies in HCI. Reflecting our own work, as we present it in the course of this thesis, the approach would have been suitable to enhance the analysis of the pointing device evaluations, such as already discussed in 2.2.3.2 (Gerken, Bieg, Dierdorf, & Reiterer, 2009a). The approach would have allowed for a much more complex study design. First, our test users could have used the mouse throughout the study and the multi-level growth curve modeling would have allowed us to compare the learning curves in detail. It would have also enabled us to identify certain peak sessions that lead to a higher amount of performance gain and compare these again between laserpointer and mouse. While the studies presented in chapters 3 regarding the PocketBee diary focus on qualitative usability results, this analysis method could help to enhance complex diary or ESM study analysis as well. We will discuss this in more detail in chapter 3.1.2.1.

2.2.4.2 Interest in Whether and When Events Occur

This research question is different from the previous ones, in that it is not concerned with the quantitative or qualitative measurement of a specific variable, but instead looks at the occurrence of an event at some point in time as a measure for change. The event occurrence by itself is the measure for change.

An example by Singer & Willet (Singer & Willett, 2003) is the research question asking the average number of years a teacher teaches at the same school. The researchers state that in general, all research questions that can be expressed as a "whether and when a certain event occurs" belong to this group. For the above research question, this would require the formulation, "Whether teachers leave their school at some point in time and if yes, when on average that happens." Accordingly, the event occurrence to look for is a teacher leaving the school. Examples from HCI research could be formulated as follows:

- "Whether and when do people sell their iPhone, compared to an HTC phone?"
- "Whether and when do people adopt a specific new technology in their daily routine?"
- "Whether and when do people rate their comfort level with a new interaction design as at least 4 on a 5-point scale?"

The last example demonstrates that many kinds of quantitative and qualitative data can be post-hoc transformed to measure event occurrence and that data for this research question is often already available.

To allow meaningful research, this research question poses two challenges. First, there has to be a substantially meaningful definition of the event occurrence. Second, there is an analysis problem when not all participants actually experience the event. This is called censoring, and it is quite common for applied research questions of this type, either because the study duration is not long enough to observe every subject encountering the event or because the event simply might not happen at all for some participants. This fact makes the analysis much more complicated than it seemed at first. Calculating an average time point for the event occurrence is suddenly not valid, as the censored data can neither be simply dismissed nor integrated with the last time-point of the study.

A possible solution, although we have not encountered it in the HCI literature, could be survival and hazard analysis methods (Singer & Willett, 2003, p. 305ff). These rather simple but powerful approaches are capable of taking cen-

soring into account and providing the researcher with a median measure of event occurrence. We describe these approaches in more detail after presenting a concrete example of these research questions from the HCI literature.

CrossTrainer: Testing the Use of Multimodal Interfaces in Situ (Hoggan & Brewster, 2010)

This study by Hoggan and Brewster is a great example of a multi-faceted longitudinal study in HCI. The researchers were interested in evaluating cross-modal audio and tactile feedback based on participant performance in the cell-phone game CrossTrainer. The study used a mixed laboratory- and field-based design. The lab-based sessions were conducted at the beginning and served as a practice baseline for all participants. Subsequently, nine participants used the game in their normal environments with time-alternating feedback variants (2 days with audio, 2 days with tactile feedback, etc.) for a total of 8 days. One of the research hypotheses, relevant here for the question of event occurrence, was that recognition performance would reach 100% in terms of recognition rate during this prolonged usage time for the different feedback mechanisms. In addition, the authors stated pre-post research questions to evaluate the level of learning for typing performance as well as interest in the average person questions, as they analyzed the typing performance for different locations visited during the field study phase.

The recognition performance question could be rephrased as whether and when participants were able to reach 100% performance. The authors did not conduct a survival analysis, instead opting to state the recognition performance at various fixed points in time. For example, they stated that participants were able to recognize the individual modality with 75% accuracy after 30 minutes of training and with 100% accuracy after 40 minutes. This is a valid and useful alternative to the survival analysis approach. However, it creates the problem of the arbitrary time-period chosen to dichotomize the individual event histories. Furthermore, it reduces the research question to the “whether” part and excludes the “when” part (Singer & Willett, 2003, p. 323). In Hoggan and Brewster’s study, this problem was not apparent, as the event (100% performance) was reached by all participants within the time frame of the study. Thus, they

were able to compute the average and the variation for when the event was reached without using more sophisticated analysis tools.

Discrete Time-Event Survival Analysis: Hazard & Survivor Functions, Median Lifetime

In the following section, we would like to introduce one specific case of event occurrence analysis in more detail, one that involves discrete time events (instead of continuous time events), and discuss the basic statistical functions that are necessary to assess the "whether and when" questions. Singer and Willett (Singer & Willett, 2003, p. 330ff) describe hazard as the "conditional probability that individual I will experience the event in time period j, given that he or she did not experience it in any earlier time period." It thus represents the risk of event occurrence for each time period. The important aspect here is that it excludes everyone who has already experienced the event. The result is a value between 0 and 1 for each data-gathering time-point; calculations for this specific case are straightforward (from (Singer & Willett, 2003, p. 332):

$$\hat{h}(t_j) = \frac{n\ events_j}{n\ at\ risk_j}$$

For example, if we have 100 participants buying an iPhone at time-point 0, then at time-point 1, all of them are at risk of selling it. The hazard function now calculates the percentage of actual sellers. For time-point 2, only those participants who still own an iPhone are still at risk – i.e., the total number of buyers (=100) minus those who sold their iPhone at time-point 1. Plotting this data, we are able to identify especially "risky" time periods and to characterize the shape of the risk development over time. Thus, we might find out that participants are more at risk of selling their iPhone 1 year after purchase than at any other point in time.

The survivor function calculates the probability that individual i will survive past time period j. It thereby calculates the percentage of individuals at time period j who have not encountered the event in comparison to all participants. The survivor function generally decreases over time and approaches zero. However, because of censoring, it seldom reaches zero. In the case in which some partic-

ipants have been censored at different points in time, the survivor function can also be calculated using the estimated hazard function.

Finally, one can now calculate the median lifetime, which describes the point in time at which the survivor function reaches .5. This is not affected by so-called uninformative censoring, that is, censoring that merely happens because we are unable to observe the event happening within the study period. Therefore, we can calculate the median lifetime (as opposed to the arithmetic mean) even though not all participants have encountered the event. This figure tells us that at this specific point in time, “half of the sample has experienced the target event, half has not” (Singer & Willett, 2003, p. 337). Thus, the average person has experienced the target event by that point.

Overall, survival analysis is a useful and easily applied tool that allows us to tell us whether and when events occur and tells us even more about the shape of event occurrence by plotting the data. As stated earlier, the data for this analysis is often already available in a study or may be obtained from defining (sensible) events post-hoc. Therefore, we encourage researchers in HCI to apply these methods in their studies, so that more information can be gathered about their utility in this specific research domain.

2.2.4.3 In-Depth Interest in the Why and How

The research questions discussed so far have taken up the challenge of investigating overall change processes, both from individual and group perspectives. In so doing, they either look for specific event occurrences or try to visualize and if possible model the change pattern over time. The advantage of these research questions is that statistical data analysis methods are available to handle the quantitative data. However, especially when exploring unprecedented situations, it might be difficult to formulate hypotheses about certain predictors of change *a priori*. In fact, it may be that in small-scale studies with a limited number of participants, we are unable to find plausible overall patterns in the data at all. Instead, we are interested in the in-depth processes, investigating not only overall patterns, but also individual changes in detail, why they occur and how they occur. Therefore, we look at the qualitative data itself instead of

using quantifications or quantitative measurements. One popular example for this research approach in HCI is Shneiderman's et al. Multi-Dimensional In-depth Long-term Case-Studies (MILCs) approach (Shneiderman & Plaisant, 2006), which we have introduced in the introduction.

An important aspect of this research question is the possibility of analyzing the data on the fly and adapting the study accordingly. This allows researchers to gain valuable in-depth insight that otherwise would have stayed hidden. This research question and the way to approach it shares some similarities with approaches such as grounded theory (Strauss, 1987), the main difference being the specific interest in change processes and the understanding of what has exactly changed, why it has changed, and how it has changed. In many cases, these kinds of studies observe and analyze their research objects over prolonged periods of time, often months or even years. It seems that in HCI these studies are often referred to as long-term studies rather than longitudinal. As with all qualitative research, the analysis of the data includes many interpretative steps. The challenges from our perspective are: 1) to "ask the right questions" to the data, i.e. focus the perspective on change processes during analysis and 2) to do this in a structured way, so that discovered changes are grounded in the data and can be referred back to the incidents that led to their "discovery." As it is often the case with qualitative research, studies may be challenged when people cannot easily follow the reasoning that led to the discoveries. A structured and replicable approach can be especially helpful here.

As with the other research questions, we will first present an example study from HCI before discussing a more elaborate approach to the analysis of such data. We will refer to the analysis framework presented by Saldaña (Saldaña, 2003) (Saldaña, 2008).

Robots in the Wild: Understanding Long-Term Use (Sung, Christensen, & Grinter, 2009)

Sung et al. present a study in which they deployed cleaning robots to 30 households and attempted to observe how the family and single households adopted this technology into their lives. They did so by means of several data-gathering

techniques, which they also had to adapt during the study, as not everything was working as expected – a circumstance that frequently occurs in longitudinal field studies. For example, they asked the participants to keep a diary, but only a few managed to actually do it. They then asked participants to send them emails whenever something interesting happened and if possible to take pictures and send them with the email.

The study was scheduled to last for 6 months, with five interviews total; one interview took place before the deployment to ask participants about their expectations and their cleaning routines. In addition, participants were asked to draw a map of their home, indicating places where they expected to use the new cleaning robot. They were asked to update this map in later interview sessions, which proved to be very helpful as a means for the participants to indicate how they actually used the robot. The second interview was the “unboxing” interview, in which participants were handed the robot and the initial experiences were captured. Two weeks later, the first “change” interview took place. As described, participants updated their maps and also created a bubble drawing to explain how they used the technology. An activity list was provided to the participants, on which they could check off several activities they might have done with the robot, such as “naming,” “watching,” or “cleaning.” The fourth interview, 2 months later, allowed the researcher to investigate in detail what had changed in the individual households and why. The same tasks were repeated again, and due to the nature of these artifacts, they were analyzed on a household level rather than on a general level (although abstractions and generalizations could be derived from that point as well). Interestingly, for some participants it seemed that the robot had become such an integral part of their daily lives that they had difficulties reporting on them. The fifth interview was scheduled after 6 months of usage, and the authors report that little changed at this point, although there was one household that only started using the robot after four months. The researchers also asked participants to do a kind of “summary” task in that meeting, in which they had to place objects in their daily life (e.g., TV, car) on a two-dimensional scale (with pleasant-unpleasant and useful-useless as the axes). They were then asked to place the robot in the final diagram,

which allowed the participants to put the robot into perspective within their daily routines.

The study is a great example of the challenges of field research in households, where it is difficult to be unobtrusive and still get the relevant and necessary information. It was also challenging for the authors to not introduce a bias, as some participants associated the robot with them and expected them to have scheduled “experiment-like” behavior during the study. When the robot happened to break, the participants thought that the researchers had programmed it to fail in order to study the effects. Although the authors had not done that, this is an interesting effect that could make longitudinal field experiments difficult to design. Unfortunately, we have not come across a design in a longitudinal field-study setup in which the manipulation of an experimental variable was done on purpose.

The authors do not report in much detail on their framework or strategies for analysis. However, given the artifacts, it is obvious that they tried to let participants make changes visually obvious, e.g., by indicating where they had used the robot on the map. Interpretation was thus much easier, and the authors were not required to ask for changes directly, which can always introduce a response bias, as participants might think that change is expected and try to come up with stories. We follow this same approach in the Concept Maps method for API usability, which we will present in detail in Chapter 4.

A Framework for Qualitative Longitudinal Data Analysis

Johnny Saldaña (Saldaña, 2003) (Saldaña, 2008) has presented a framework that tries to cope with the challenge of how to analyze and interpret qualitative data for changes. This is especially difficult if one cannot rely on artifacts, as in the study by Sung et al. (Sung, Christensen, & Grinter, 2009), or in the Concept Maps method (Gerken J. , Jetter, Zöllner, Mader, & Reiterer, 2011) that make such changes accessible, but that relies mainly on text notes from field observations and interviews. Saldaña addresses the challenge of making the analysis process reproducible to some extent and of giving it a clear structure and format so that others can easily relate to it. He puts the main focus on the coding of the

data, providing a template/matrix that is meant to be used for “one period of time” (see Table 3). This does not necessarily have to be a single interview session but can also be on a more cumulative level, if data was gathered more quickly than things were changing and there might be little added knowledge to code the data on the lowest possible level. Saldaña identified 7 indicators for change processes, and each of them is represented in one column:

- Increase/Emerge:
  The focus here is on new events that have not happened before. Saldaña asks questions such as what increases or emerges over time: “an increase or emergence in qualitative change is a phenomenon or participant action that appears or transforms in subtle, smooth, or expected ways” (Saldaña, 2008).

- Cumulative:
  Here, Saldaña looks for aspects that only add up to a more complex, cumulative phenomenon over time, with at least three matrix pages (or time periods) leading up to it. Saldaña calls this “three-cell time triangulation,” pointing out that cumulative change is not always a smooth path but can only become obvious after several data-gathering periods analyzed from an overall perspective.

- Surge/Epiphany/Turning Point:
  Here, he asks the question of what kind of surges, epiphanies, or turning points occur over time. This describes major events, experiences, or personal revelations of a magnitude that significantly alters the things under observations (e.g., attitudes, values, belief systems).

- Decrease/Cease:
  Change can go in both directions, so it is important to also ask for what is decreasing or even ceasing over time. The danger is that the researcher may change the way he or she records the data over time, and that aspects

that seem to be decreasing when analyzing the data have simply no longer been recorded in detail (e.g., because they became routine).

- Constant/Consistency:
  This leads us to the next aspect – one should also examine and explicitly record what stays constant over time: the things that have not changed and that remain stable. This is often actually the largest part of the observational data, and researchers must take care to not forget about this aspect when looking for changes. Knowing what remains constant provides the perspective to interpret changes, or as Saldaña notes, "We cannot discern what is changing unless we also know what is not changing" (Saldaña, 2008).

- Idiosyncratic:
  Here, Saldaña suggests marking the special things, the anomalies in daily routines. These are different from turning points, as they are rather "inconsistent, ever-shifting, multidirectional and, during fieldwork, unpredictable" elements in the data.

- Missing:
  The final aspect asks the researcher to look for missing things. Thereby, Saldaña means everything that would have been expected by the researcher and simply is not present in the data: "We note that [phenomena that are] most possibly and plausibly missing as they relate to what is present" (Saldaña, 2008). We think that this is a very thought-provoking idea, as it constantly asks the researcher to think about the data and not just report it. The interpretation is already triggered at that point, helping the researcher to reflect on the data.

While this first process is rather descriptive, the next steps now ask the researcher to find differences between the individual matrix pages. Therefore, Saldaña provides several additional rows and columns in his matrix template. This allows the identification of changes to become a systematic process, while

still being interpretative and subjective. It allows the researcher and others to trace back the interpretations. We will present the most important ones below.

- Differences above from previous data summaries:
  Here, the researcher marks and highlights everything which is different from previous observations. He or she does so not only for an individual cell (e.g., the increase cell) but across all the dimensions noted above. Saldaña describes this as “active thinking” about differences and a search to find everything that stands out.

- Contextual/intervening conditions influencing/ affecting changes:
  This asks the researcher to identify how, how much, in what way, and why the observations above have occurred. The “influence and affects” is the replacement for “cause and effect” from quantitative research, and the basic idea is to look for and identify the conditions that led to the data as it is.

- Interrelationships:
  Here, Saldaña asks for changes that interrelate through time. He acknowledges that this is a highly interpretative step and in principle one can relate everything to everything else, which makes it difficult to identify the really important correlations in the data. He also notes that one always should look for additional data to confirm these relationships, e.g., the researcher should specifically ask participants in subsequent interviews whether the interpretations done in this step are correct.

- Changes that oppose/harmonize with human development/social processes:
  This row asks the researcher to think about whether or not the findings fit the theory or other empirical research. One must keep in mind that this is still about changes and whether or not changes happen as expected.

- Preliminary assertions as data analysis progresses:
  This is, as Saldaña calls it, a “think out loud” row and seems to be very similar to the “memo” process in grounded theory. It asks the researcher to cre-

ate memos about his thinking and to simply mark everything that might be interesting for the analysis. Saldaña states that “whatever works” should be the paradigm of thinking here and that “the longitudinal qualitative researcher’s analytic process is neither completely linear nor holistic. It is iteratively – if not erratically – cumulative and serendipitous in knowledge building” (Saldaña, 2003).

- Through Line:
  The “through line” describes the ongoing story of the study; it is the essence of this matrix page and a bit similar to a conclusion. “The through-line helps navigate the researcher’s journey as he or she writes the final epic of his participants’ changes (if any) through time” (Saldaña, 2008). This through line should capture the story of changes as one would start reporting it in a paper.

Saldaña, based on his background in theatre research and the arts, claims that research is like watching a play in which the audience constantly asks “what happens next.” He states that this is also the driving question for longitudinal observation fieldwork and for analysis. It is important to note that he warns that this approach is not meant to be used for hundreds of participants, as it is obviously quite time-consuming. He also states that analysis can influence further data-gathering and in some areas it should (e.g., the interrelationships). To our knowledge, his approach has never been used in HCI, but could well be applied and adapted. Especially valuable are the themes or indicators for change, which help the researcher to ask the right questions of the data. As Saldaña states, the approach is flexible and can be done in collaboration with other researchers as well. It could also be shortened or extended to fit the needs of the study. Overall, we think it could be a very good starting point for a similar framework for qualitative longitudinal field studies in HCI.

Table 3: Template/Matrix by Saldaña to analyze qualitative change over time (Saldaña, Analyzing longitudinal qualitative observational data, 2008)

Increase/ Emerge	Cumulative	Surge/ Turning Point	Decrease/ Cease	Constant/ Consistent	Idiosycnratic	Missing
	Differences above from previous data summaries					
	Contexutal/intervening conditions influencing/affecting changes above					
Interrelationships						
Changes that oppose/harmonize with human development/social processes			Participant/Conceptual Rhythms			
Preliminary Assertions/Memos			Through Line			

2.3 A Taxonomy for Research Designs in Longitudinal Research in HCI

The previous chapter provided a comprehensive overview of possible research questions in longitudinal HCI research. To some extent, this included a discussion of several different research designs, with the longitudinal panel design predominating. However, while this is generally one of the most suitable designs to conduct longitudinal research, there are several alternatives. In this chapter, we will present a taxonomy of the different possible research designs. The driving principle here again is to derive a taxonomy that reveals (hierarchical) relationships between study designs that share fundamentally similar structures. In addition, we will explain how they are related to the research questions, data-gathering schedules and methods. We will start by discussing two general aspects of longitudinal research designs: the study duration and the data-gathering schedule.

2.3.1 Study Duration

An important aspect of any longitudinal design is the notion of time. As discussed in the research question section, time is not only an organizational characteristic for a study but also an intra-individual factor that may determine change processes. Therefore, a typical question during the design of a longitudinal study is how long the study should actually last. Longitudinal studies have often been understood as studies that take place over a very long period of time, e.g., over several months or even years. For example, for qualitative longitudinal research in the educational sector, Saldana (Saldaña, Longitudinal Qualitative Research: Analyzing Change Through Time, 2003) suggests studies of at least 9 months; in the social sciences in particular, there have been studies that analyzed census data over several decades. However, time can also be expressed in terms of how many data-gathering waves take place. Here, both Singer and Willett (Singer & Willett, 2003) and Karaponas et al. (Karapanos, Martens, & Hassenzahl, 2009) suggest that at least three data-gathering waves are required to be able to capture changes. While a longer duration and more

data-gathering waves should generally lead to a more comprehensive view of the change processes, one should not disregard the cost associated with such a study and whether it is actually economically effective for its purpose. In addition, data-gathering waves often disrupt the natural environment of participants, both in field and lab studies, and too many waves can introduce a bias in the measurements. From our perspective, the answer to the question of a study's duration should always depend on the specific research question at hand and on the kind of changes to be tracked over time. It might well be that change processes can be sufficiently observed within a 2.5-hour session with several measurement waves, in case the changes are caused by low-level motor skill learning processes that can be achieved with intensive training within this time period, such as in (Rieger, 2009). Observing how people change their attitudes towards a product they have purchased, however, might require a longer study and one that potentially reveals many ups and downs. In the end, one must select a data-gathering schedule that can track the change frequency without missing ups and downs and a sensibly chosen study duration that allows observation of the changes actually taking place. Consequently, it is important that researchers reporting on the results of a longitudinal study state arguments in support of their specific schedule and study duration and illustrate the possible limitations thereof.

2.3.2 Equal vs. Unequal Data-Gathering Intervals

Figure 13: Equal time intervals among four waves of data gathering (blue bars) across time (t1-t4)

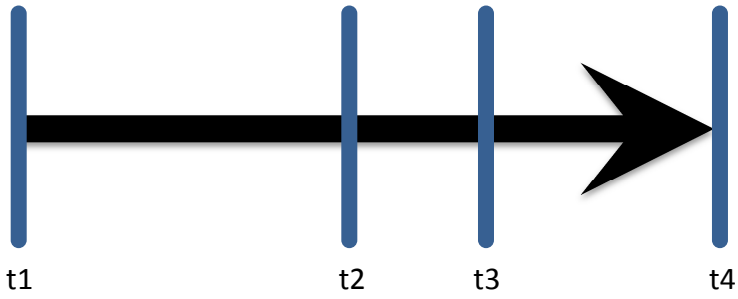


Figure 14: Unequal time intervals among four waves of data gathering (blue bars) across time (t1-t4)

When considering the data-gathering schedule, another question (in addition to the number of intervals) quickly arises. When there are more than two points in time for data-gathering, do we have to schedule equal measurement intervals? And is it important that every participant encounters the same number and schedule of intervals? Again, there are several answers to these questions. First, it depends on the coding of the time variable. As mentioned earlier, in principle we can simply code a "real" time variable, such as the age of participants or the number of days they have been using a product. Another possibility is to code the data-gathering waves directly as measures of time. The first option, using a "real" time variable, is the predominant approach in the social sciences and should be preferable in cases, where the researcher does not directly control what happens in between data-gathering waves. This is the case for almost all field studies. In addition, this type of time variable is easier to interpret as any change processes can be directly related to a common notion of time (e.g. it took our participants 5 days to learn the system). The second option, using the data-gathering waves to code time, is common in experimental setups, probably as these setups are often designed in a manner similar to traditional cross-sectional repeated-measures experiments. Such an approach is often very convenient, as the data-gathering waves need no further coding other than the wave number. However, it would obviously be dangerous to design this type of study using unequal intervals (i.e., different time periods between the intervals), as this information would be lost during the coding process. However, for short-term lab-based longitudinal experiments such as those in (Gerken, Bieg, Dierdorf, & Reiterer, 2009a) or (MacKenzie & Zhang, 1999), this

is a sensible approach and can prevent “over-interpretation” of the time variable. For example, in (Gerken, Bieg, Dierdorf, & Reiterer, 2009a) we scheduled five sessions of 30 minutes each of laser-pointer use on five consecutive days for each participant. Using a real-time variable, one would conclude that a certain level of performance was reached after 5*30 = 150 minutes of usage (or five days of usage at 30 minutes per day). However, this neglects the fact that we do not know the influence of the interval duration of one day between the data-gathering sessions; performance could perhaps appear quite different with longer or shorter intervals of abstention from the activity.

Using real time can offer other benefits, especially for long-term field studies. Given adequate analysis techniques, such as the multi-level growth modeling we have presented, we can incorporate not only unequal intervals but also different intervals for every participant. This makes longitudinal study design much more flexible and reduces the organizational overhead for both the researcher and the participants, who can then follow much less obtrusive schedules.

Finally, data-gathering techniques such as logging or (to some extent) diaries (Bolger, Davis, & Rafaeli, 2003) allow for continuous data-gathering that does not require any pre-defined schedules. Here, the coding of a real-time variable is especially important, as a post-definition of “waves” would result in a loss of data.

2.3.3 Panel Designs

The basic idea of all panel designs is to follow the same group of participants over several data-gathering waves. It is thus the most natural way of conducting longitudinal research, as it allows us to look for inter- as well as intra-personal changes over time.

2.3.3.1 Within-Subjects Repeated Sampling

Figure 15: A within-subjects repeated sampling design with two data-gathering waves – at the beginning and end oft he study.

The basic principle of within-subjects repeated sampling, as discussed by Karaponas et al. is the existence of only two data-gathering waves, thus limiting the possible research questions to pre-post comparisons (Karapanos, Martens, & Hassenzahl, 2009). For this type of research questions we already discussed the study by Kjeldskov et al. (Kjeldskov, Skov, & Stage, 2005), “Does time heal?” in which the authors studied a hospital patient record system and compared usability problems at participants’ introduction to the new system and 18 months later. Karaponas et al. (Karapanos, Martens, & Hassenzahl, 2009) explicitly distinguish within-subjects repeated sampling from panel designs, describing the latter as consisting of at least three data-gathering waves. However, from a methodological point of view, within-subjects repeated sampling is still a panel design, since the same group of participants is followed over time. Therefore, we have decided to classify within-subjects repeated sampling as a specific form of panel design rather than its own category.

2.3.3.2 Prospective Panel Design

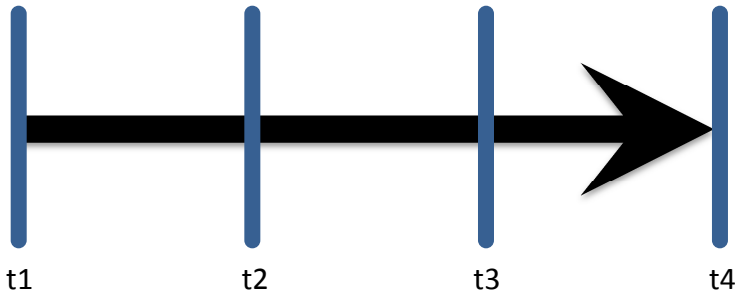


Figure 16: A prospective panel design with four data gathering sessions and equal intervals

Prospective panel designs, as opposed to within-subjects repeated sampling, incorporate at least three data-gathering waves, allowing research questions interested in the process of change. Several authors (e.g. (Singer & Willett, 2003) (Karapanos, Martens, & Hassenzahl, 2009)) argue, and we concur, that it is not possible to analyze change processes in detail with only two data-gathering waves, as this would diminish any complex change process to a simple “up” or “down” conclusion.

A prospective longitudinal panel, while the most natural approach to longitudinal research, also incorporates many of the challenges discussed in Chapter 1. More specifically, data gathering is inherently interdependent and measurement constructs may become invalid over time, as participants change or they simply become accustomed to the measurement in unwanted ways. For prolonged periods, participant drop-outs (panel attrition) and study organization can also develop into severe problems. In addition, it is difficult to account for external factors that might or might not influence participants over time. Imagine a study analyzing the use of a mobile device operated by a pen in a longitudinal panel design. If the release of the iPhone, as a device that made touch-based interaction from one day to the other the state-of-the art, falls right in the middle of this study, there might be good reason to assume that this event could have an effect on the participants of the study and how they perceive pen-based interaction.

2.3.3.3 Revolving Panel Design

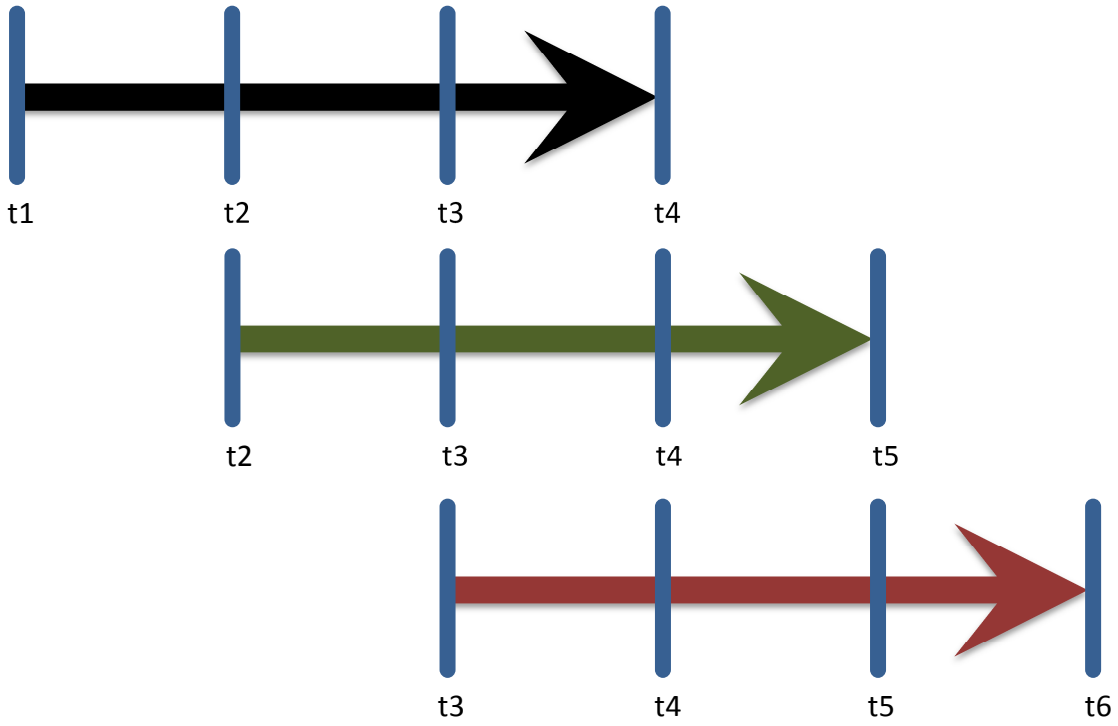


Figure 17: A revolving panel design with new waves of participants joining at each data-gathering wave. Each arrow represents a different set of participants.

A solution to the problem of the influence of external factors as well as to reduce the effect of panel conditioning (as discussed in chapter 1) might be found in the revolving panel design (Menard, 2008). While we have not seen evidence of this method's use in HCI, it is a promising approach to resolving some of the typical problems of panel designs. A revolving panel includes the introduction of new participants or research units during the course of the study. For example, at each wave of data gathering, existing participants are joined by a (probably smaller) number of new participants. Thereby, it is possible to compare the effects one would ascribe to changes in behavior with an unbiased group. Thereby it is possible to test whether external factors also play a role in the observed effect, or whether the measurement itself has created a certain response bias. In order to avoid being overloaded with too many participants at the end of the study, the design should also include release phases for participants after a certain period of participation.

2.3.3.4 Longitudinal Case Study

A special case of prospective panel design is the longitudinal case study. The idea of this design is to observe only a small number of participants but over a long time period, e.g., for several months or even years. Researchers are meant to analyze these subjects in detail without the need to derive generalizable results. The aim is to fully understand their behavior or their interaction with a product in question. Shneiderman and Plaisant (Shneiderman & Plaisant, 2006) promote such an approach with their MILC (Multidimensional In-depth Long-term Case Study), which we have already introduced in the introduction.

2.3.4 Repeated Cross-Sectional Designs

The primary element that distinguishes this method from panel designs is the use of different participants at each data-gathering wave. We can further distinguish between designs that incorporate repetition over time and repetition over different user groups.

2.3.4.1 Repetition over Time

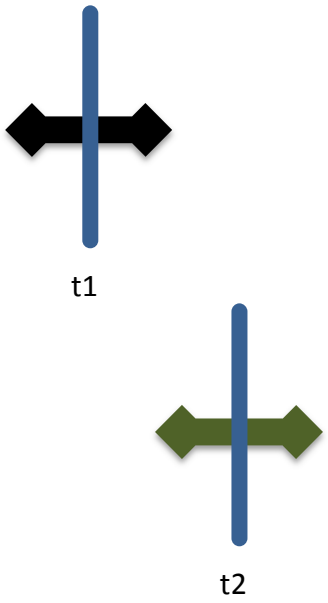


Figure 18: A repeated cross-sectional design, with two distinct cross-sectional studies at t1 and t2 (different user groups)

For this type of design, a different sample of participants is recruited for each data-gathering wave. However, the samples should be comparable for certain study-relevant attributes, as they will be analyzed following the same procedure as in longitudinal studies. As described in the introduction, these designs are necessary if institutional/extra-individual change is the subject of study. For example, studying whether the working conditions in coal power stations have improved over the last 30 years would require such a design. A longitudinal panel in contrast would be 1) difficult to implement, as not many participants are available who have been working in the same coal power station 30 years ago and today and 2) introduce intra-individual change processes as well, which may bias the institutional change process. Besides, the approach offers the advantage that different cross-sectional studies can be combined into a longitudinal design, in some cases even after the fact. For example, in the social sciences it is quite common to analyze *per se* cross-sectional data, collected in regular waves and with well-defined sampling techniques, combined into a set of longitudinal data. As different user groups are studied, there is no possibility to investigate intra-individual change processes. A particular danger lies within low sample sizes (as often common in HCI) as they increase the danger of mistakenly attributing differences between groups to changes over time, although they may be caused by the sampling groups being themselves very different.

Still, especially from an organizational point of view, such designs provide clear advantages, as every study could provide answers to cross-sectional research questions in addition to the longitudinal research question.

2.3.4.2 Repetition over Different User Groups

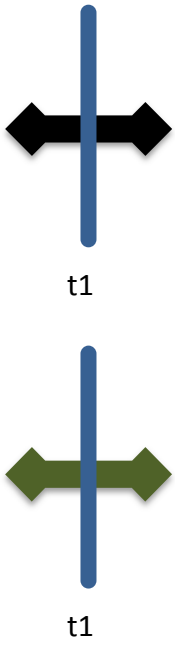


Figure 19: Repetition over different user groups to study the impact of anticipated change processes

A “short-cut” to longitudinal research in HCI is the use of different user groups in a cross-sectional study, with the assumption that the differences between users resemble differences over time. For example, to analyze how people learn an interface, a common approach is to invite both novices and expert users. The difference between the groups is then attributed to the learning time the expert users have had on the interface. While this “discount” approach may offer some appeal, there are certain inherent problems. First, as with repetition over time, it is impossible to analyze intra-individual change processes. Second, it is probable that at least some of the differences between such user groups are not due to the “time-dependent” variable of experience/learning but rather to individual differences. Third, such a “time-dependent” variable is often a huge simplification, as has been stated several times in the literature (Kjeldskov, Skov, & Stage, 2005) (Karapanos, Zimmermann, Forlizzi, & Martens, 2009). To become an expert user, time is only one factor for possible differences. Training (explicitly or as a product of time), dedication, and individual skills are at least equally important. Thus, to be able to attribute any differences in the measurements to

the experience level, one must make sure that the novice group is at least equal in terms of individual skills and dedication.

2.3.5 Retrospective Panel Designs

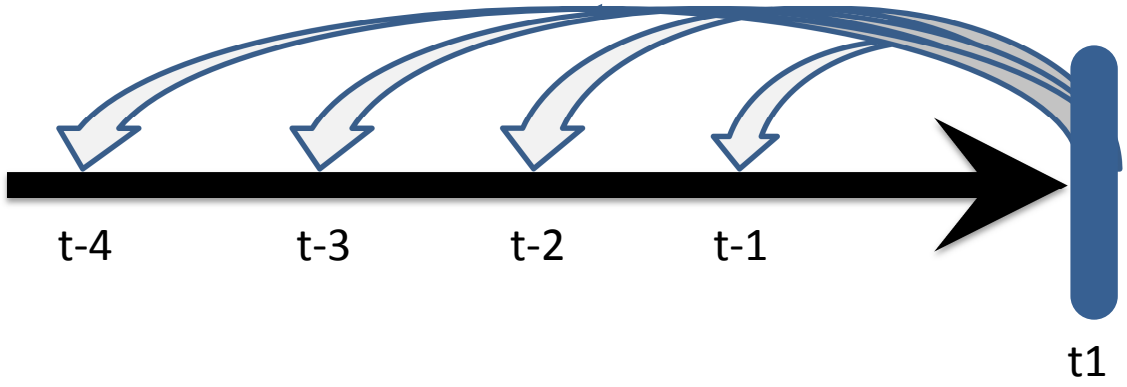


Figure 20: Retrospective Panel Design which „looks back in time“

Menard describes a retrospective panel design as being “identical to a prospective panel design in every respect except the number of times data collection actually takes place and the length of the recall period required of respondents” (Menard, 2008). The first aspect, the number of times data collection takes place, refers to the fact that a retrospective panel design collects data only once, but does so retrospectively for two or more periods in the past. The second aspect, the length of the recall period, refers to the fact that in principle every interview/survey that asks a participant to report on certain events is inherently retrospective. However, a retrospective panel design is fundamentally different, as it systematically asks participants to report on multiple past events that they experienced or observed at different distinct points in time.

Retrospective panel designs offer a variety of advantages over prospective panel designs. First, there is no danger of panel attrition, as participants either take part in the study or do not. Second, the costs of such a study are much more predictable and attractively lower, again because only one data-gathering wave is needed. Third, as Karaponas et al. argue in certain cases experiences reported from memory can be of more interest than the actual experience itself,

as “these memories (1) will guide future behavior of the individual and (2) will be communicated to others” (Karapanos, Martens, & Hassenzahl, 2009).

However, there are also certain drawbacks to consider. First and foremost, retrospective panel designs can only rely on indirect data-gathering techniques, such as inquiry or reconstruction techniques, but obviously not on direct observation or measurement of performance at different points in time. This focuses the value of retrospective designs toward research questions that show interest in opinions, feelings, and the reporting of events or situations. Considering that we are interested in analyzing the learning process for a new interactive system, we would have to rely on participants’ reports of their feelings, the usability issues encountered, etc., and how these changed over time, but we would be unable to actually measure any of these directly. Second, such designs introduce a memory bias, as participants must report on events that often occurred weeks, months, or even years in the past. Menard (Menard, 2002, p. 44) states that such designs work better with salient events than with attitudes or other psychological data, as events are more objective. In case of attitudes, participants tend to create a consistent life story and align earlier attitudes with their current ones. Through the act of retrospection, they reflect upon their life or the period in question and reinterpret their memories. Whenever possible, short term retrospection should be preferred. Nevertheless, Taris states that “a prospective longitudinal design will virtually always result in better (more reliable and more accurate) data than a retrospective design” (Taris, 2000).

Considerable effort has been conducted to improve this memory bias, such as the iScale approach (Karapanos, Martens, & Hassenzahl, 2009), the Day Reconstruction Method (Kahnemann, Krueger, Schkade, Schwarz, & Stone, 2004), and the CORPUS interview technique (von Wilamowitz-Moellendorf, Hassenzahl, & Platz, 2007). All of these techniques involve a certain explicit construction process which is meant to help participants to remember their actual experiences and correctly place them in the time dimension.

iScale for example aims at eliciting the longitudinal user experience by asking the users to sketch these on a simple time-scale with a UX dimension on the y-

scale. Important events could be marked and annotated and the authors even developed an electronic tool that implements this technique. Figure 21 shows an example of such a drawing with the ease of use on the y-axis and time on the x-axis. The numbers and annotations mark certain events in time and allow the researchers to understand these events and their relationship with the overall usage experience.

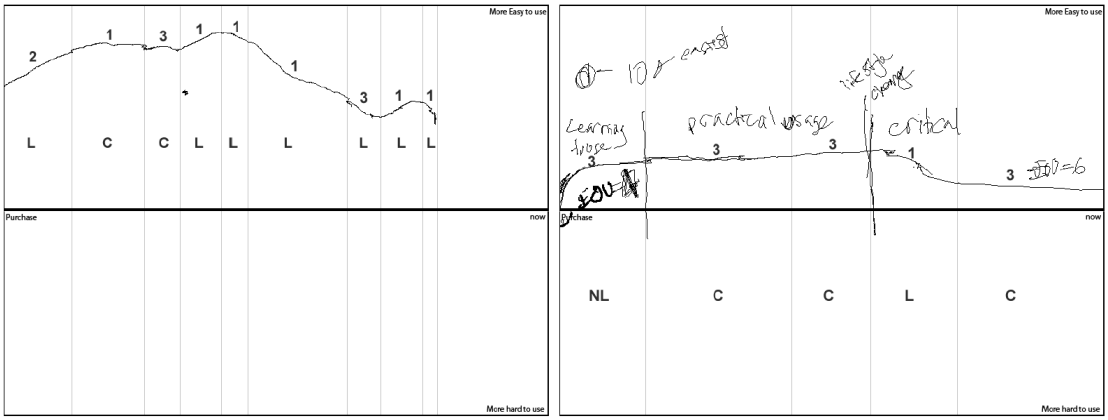


Figure 21: Example graphs from a iScale study with hand-sketched graphs (taken from (Karapanos, Martens, & Hassenzahl, 2009)).

Another often overlooked drawback pointed out by Menard is the inherent bias of participant selection, or sampling bias, as he calls it (Menard, 2002). In the most extreme case, participants who have died before the retrospective study takes place are completely excluded from taking part in such a study, even though they could have been part of a prospective panel design for the first few data-gathering waves. While this situation is rather uncommon in HCI research, the problem still applies to a lesser extent in other situations. Consider a retrospective study that tries to identify how the experiences of using a certain mobile phone change over time. Naturally, one would try to find participants who are owners of this mobile phone; however, this would implicitly exclude all those who have stopped using the device and e.g. changed to a competing product. In this specific case, it is possible to also search for participants who have owned the specific device in the past to avoid this bias. However, it is important to think about the kind of participants that are naturally excluded from the study if it is mistakenly framed in this intuitive but incorrect fashion.

2.3.6 Relationship between Research Questions and Research Design

While there is no easy one-to-one mapping between individual research questions and research designs, some designs are more suitable for certain types of research questions than others, and some may be simply inappropriate (even if technically possible). In Table 4 we give an overview of this relationship. The green fields mark the most appropriate research designs for a certain research question type, while the red fields mark the designs that are, from our perspective, not well-suited to the question type. The fields that are left blank are also possible research designs and for certain particular questions might even be the best choice.

For researchers conducting a longitudinal study to obtain more data about individuals to analyze averages or cumulative data over time, a prospective panel would be the best way to proceed. A revolving panel, while possible, would not offer any benefits, as the data is not analyzed for changes. A retrospective panel design is possible but in most cases the inherit drawbacks, such as memory bias make this design less powerful. Researchers interested in pre-post comparisons can often rely on within-subjects repeated sampling with only two data-gathering waves (as in (Kjeldskov, Skov, & Stage, 2005)), but again other designs are possible. Repeated cross sectional designs are not suitable to investigate intra-individual change processes and therefore are most of the time not appropriate for studies that investigate event occurrences or the in-depth why and what of change processes. The examples we have presented for interest in the shape of change are also most of the time interested in individual change processes, again making the designs inappropriate. In general, we would not recommend the use of repeated cross-sectional designs if not for the reasons stated above: the interest in changes in an institutional or extra-individual variable. Retrospective designs are not well-suited for event occurrence research questions, as discussed above, since the data-gathering too easily misses participants who have been censored by the time the study takes place.

Table 4: Relationship between research questions and research designs

	Panel Designs				Repeated Cross-Sectional Designs		Retrospective Panel Designs
	Within-subjects repeated sampling	Prospective Panel	Longitudinal Case- Study	Revolving Panel	Repetition over time	Repetition over different groups	
Averages/Cumulative	✓	✓	✓	✗	✓	✗	✓
Outcome	✓	✓	✓	✗	✓	✓	✓
Pre-Post	✓		✓	✗	✓	✓	✓
Shape	✓	✓	✓	✓	✗	✓	✓
Event occurrence	✓	✓	✓	✓	✗	✗	✗
What & Why	✓	✓	✓	✓	✗	✗	✓

Overall, in most cases of longitudinal research, a prospective panel design would be the optimal choice, increasing also the flexibility for post-hoc decisions on data-analysis. Then again, it is the most complex and costly design and for certain questions, cheaper alternatives may be preferred.

2.3.7 Data-gathering schedules

As explained in the previous section on retrospective designs, there are different possible data-gathering schedules in longitudinal research. Table 5 shows the relationship between research designs and the various data-gathering schedules. A **continuous data-gathering** schedule basically records a continuous stream of data. For purposes of analysis, the researcher may dichotomize this stream into distinct waves. The advantage of such continuous data-gathering is that no data is lost in between the waves. The difficulty is 1) to actually gather data in this way and 2) to analyze the data without getting lost in the large amount of "noise." In principle, only automatic electronic logging or video capturing is capable of gathering data in this way. However, one must be aware that logging loses a great amount of context information that might be valuable to the research question at hand. Triangulation of data-gathering sources is of at least equal importance in longitudinal research as it is in cross-sectional studies, especially in combination with logging techniques (Gerken J. ,

Bak, Jetter, Klinkhammer, & Reiterer, 2008a). To some extent, diaries can also provide an approximation to continuous data-gathering, in case participants create reports on every relevant event by themselves.

In most cases, while there might be some continuous data available, there will also be data captured at certain **pre-defined intervals**. As discussed above, apart from pre-post comparisons in within-subjects repeated sampling designs (and their equivalent of repeated cross-sectional designs with repetition over time), we should rely on more than two data-gathering waves. Finally, retrospective designs as well as cross-sectional designs that include repetition over different user groups rely on a **single data-gathering wave**. While retrospective designs still capture data for multiple points in time, cross-sectional designs include the time variable by a systematic variation of the user group within the study.

Table 5: Relationship between research designs and data-gathering schedules

	Panel Designs				Repeated Cross-Sectional Designs		Retrospective Panel Designs
	Within-subjects repeated sampling	Prospective Panel	Longitudinal Case- Study	Revolving Panel	Repetition over time	Repetition over different groups	
Continuous	✗	✓	✓	✓	✗	✗	✗
Multiple waves (>2)	✗	✓	✓	✓	✓	✗	✗
Once	✗	✗	✗	✗	✗	✓	✓
Twice	✓	✗	✗	✗	✓	✗	✗

2.3.8 Data-Gathering Techniques and Methods

Regarding data-gathering techniques, methods, and approaches, we can identify two main challenges: 1) whether the technique can produce valid measures when being applied multiple times (construct validity), and 2) whether the method collects data in such a way that it allows analysis of changes. In principle, every cross-sectional data-gathering method and measurement instrument can be applied within a longitudinal design as well; however, not all of them are

equally well-suited to coping with these two challenges. With respect to the first problem, a simple but easy to comprehend example is the use of an IQ-test instrument in a longitudinal study design that investigates IQ development from infancy to childhood. The problem here is that it is simply impossible to use the same test instrument for both infants and for young children. Instead, the researcher has to use an appropriate test instrument for each phase of the children's development that should still measure the same construct. However, insuring that this is the case requires extensive pre-testing of the data-gathering technique. Singer & Willett refer to Lord (1963) with regard to this issue as follows: "just because a measurement was valid on one occasion, it would not necessarily remain so on all subsequent occasions even when administered to the same individuals under the same conditions" (Singer & Willett, 2003, p. 14). As an example, they take a multiplication test, which may be a valid instrument for mathematical skill, but may become a measure of memory when administered multiple times. Even "objective" measures can be biased over time. Consider again the studies of Gerken et al. (Gerken, Bieg, Dierdorf, & Reiterer, 2009a) and Bieg (Bieg, 2008). While participants improved up to 2 bits/s during the practice task, only a fraction of this gain transferred to the actual test task, although that task was only slightly different. Here, it seems that the practice task increased a "learning bias," such that participants not only learned how to use the laser-pointer, but also the specific task at hand. As a result, the performance measure was no longer a sole performance measure for the laser-pointer, but rather for the combination of laser-pointer and the specific task. While this is obviously often the case in experimental studies, longitudinal designs can amplify such effects even further, leading to false conclusions if the researcher does not incorporate appropriate control mechanisms (such as a different test task). Regarding qualitative data-gathering techniques, apart from observation techniques of which participants are unaware, every approach will itself have an impact on the study design, similar to Heisenberg's Uncertainty Principle in quantum physics (Saraiya, North, Lam, & Duca, 2006). An interviewer asking a participant about changes may invoke a reflective phase within the participant; going forward, this might impact the next data-gathering wave as the participant expects to be asked for changes again. As a result of such

panel conditioning (Cantor, 2008), changes may be overrepresented in the data as participants start to look for them. On the other hand, if the researcher does not explicitly ask for changes, then some may be overlooked, as participants may simply forget about them. In the end, there is no easy solution to this issue. An approach that we have followed in our Concept Maps approach (presented later in detail) is the idea of using constructive activities as means of data-gathering, and the modification of these as means of change measures. Thereby, participants are not directly asked for changes, which might help to reduce their oversensitivity to changes. Such techniques have also become very popular in retrospective panel designs for similar reasons, with the aforementioned iScale or Day Reconstruction Methods (Karapanos, Martens, & Hassenzahl, 2009), (Kahnemann, Krueger, Schkade, Schwarz, & Stone, 2004). Moreover, these techniques also implicitly address the second challenge, allowing the researcher to actually identify changes. If quantitative measures are available, changes can be easily detected by the use of statistics. However, the situation is often more difficult for qualitative data, as was discussed in the research questions section. Constructive methods such as the Concept Maps approach can make changes visible and traceable over time that otherwise would have to be inferred from text or audio-visual data. In this way, we can in essence avoid an additional interpretative step in the analytic process, also facilitating analysis of such data in teams.

In addition to these methodological challenges, a practical challenge often arises: how to actually gather data multiple times while still maintaining sensible levels of effort and expense? Apart from the typical cross-sectional data-gathering techniques, longitudinal studies have led to the development of two specific approaches with the purpose of gathering data over time with the lowest possible effort by the researcher. First, logging approaches allow the researcher to gather interaction data without any obtrusive measurement influence – and in most cases, remotely and in real time (Lazar, Feng, & Hochheiser, 2009, p. 307ff). Second, diary and ESM approaches ask the participant to maintain a data-gathering record and thereby also reduce the costs for the researcher and often any retrospective bias of survey or interview techniques. From our perspective, both of these approaches offer great potential for

improved tools and techniques to support the researcher. With PocketBee, we have focused on the diary and ESM approaches and will present in detail the concept of such a multi-modal diary for field studies in Chapter 3.

2.4 Implications & Conclusion

In this chapter, we have presented and discussed a taxonomy for longitudinal research in HCI. We have focused on two central aspects here – the research questions that should drive every study and the research designs. Besides, we have illustrated the relationships between these two along with data-gathering schedules and methods as well as analysis techniques.

Beyond the taxonomy itself, this chapter serves as a comprehensive introduction to longitudinal research in HCI and related disciplines, providing a thorough discussion of challenges and benefits. We have carefully selected example studies, mainly from HCI to illustrate the main aspects. Furthermore, as the taxonomy was in large parts derived based on a literature review of existing studies, we provide a classification table of all reviewed studies and how they relate to the taxonomy. This can be found in Appendix A.

For the purposes of this thesis, the taxonomy, while being constantly revised and extended over the years, served as the foundation for the identification of worthwhile research areas within the field of longitudinal research. As we have discussed throughout the thesis so far, there are still many open challenges, e.g. regarding problems such as panel conditioning or attrition. For the remainder of the thesis, we have focused on two more practical areas.

First, our research so far has shown, that longitudinal research often requires much more effort and costs on the side of the researcher. Therefore, we think it is essential to put more effort into the design and development of tools which support the researcher and allow a more convenient and also more powerful and flexible data-gathering design. The PocketBee diary/ESM tool was designed with this challenge in mind and will be presented in detail in the next chapter, alongside a thorough discussion of the diary method per se.

Second, we have identified that especially qualitative longitudinal research suffers from increased challenges regarding the data analysis. The approach we presented by Saldaña, while being promising, has not yet been applied in HCI. The main problem is that qualitative data does not easily show change processes as with quantitative data. Therefore, the researcher does not only have to put more effort in the analysis, it is also more difficult to design appropriate data-gathering methods which actually capture the change processes. In chapter 4 we will address this issue by presenting a constructive approach to visualize changes in the mental model of a user over time. As a usage scenario we have picket the evaluation of application programming interfaces (API). These are particular difficult to evaluate with “standard” cross-sectional usability evaluation methods and could benefit a lot from longitudinal studies, as these would then allow to study learning processes over time.

3 Using Diaries for Longitudinal Field Research in HCI

One of the biggest challenges and also cost factors in longitudinal research is the necessity to organize and conduct multiple data-gathering sessions over the course of a study. While the organizational matters are not negligible, this also severely increases the external influence and bias effects the researcher may create for participants. Data-gathering itself is an intrusive act and thereby can cause unwanted reactivity in those being observed, such as the Hawthorne effect or a general observer effect (Bolger, Davis, & Rafaeli, 2003). This is especially true for methods that require the researcher to interact with the environment, such as conducting interviews or participatory observations; in a longitudinal setting, this simply multiplies the chances of such effects.

Remote research methods offer an intriguing solution to such issues, as they allow data gathering in its “natural, spontaneous context” (Barrett & Barrett, 2001) without being obtrusive and thereby can be used in situations in which observation or experimentation would be impossible or inappropriate. As data-gathering happens *in situ* and without the need for explicit observation or interview sessions, techniques and methods such as logging (Lazar, Feng, & Hochheiser, 2009), diaries (Bolger, Davis, & Rafaeli, 2003), or ESM (Hektner, Schmidt, & Czikszentmihalyi, 2007), are especially suited for longitudinal research. In HCI, interaction logging has enjoyed an increased popularity and is already an established part of e-commerce analysis processes. Services such as Google Analytics have become very popular but also raise questions of privacy invasion. However, as has been documented by HCI researchers several times, e.g. (Gerken J. , Bak, Jetter, Klinkhammer, & Reiterer, 2008a), logging as a data-gathering technique suffers more than any other technique from lack of context; this makes gathering data beyond descriptive usage patterns a difficult and often impossible task. Diaries and ESM address this problem of data-gathering in the wild, so to speak, from the completely opposite direction. Instead of providing tools and techniques for researchers to gather the data au-

tomatically, they ask the participants to do the data-gathering. This has several striking advantages, including reduced retrospective bias compared to interview techniques and the possibility to gather data as “it happens”. As discussed in Chapter 2, we find the diary method especially compelling, since it offers a great deal of flexibility regarding its application, with the possibility to gather both quantitative and qualitative data. We will further show in subsequent sections that diaries and ESM share many similarities and should be regarded under the same methodological umbrella. Diaries in longitudinal research design can also be easily combined with almost any other data-gathering technique – logging, for example, or interviews using the diary entries as prompts to allow in-depth discussions. In section 3.3 we present a study which illustrates this triangulation of data-gathering methods in the context of diaries.

In general, a diary is foremost a private record of a person that may include facts as well as subjective judgments or personal stories. It is kept on a regular basis (e.g., every day or for every important event) and discusses or describes contemporary events. Thus, it is inherently a longitudinal data-gathering method. When applied as a research method, the diarist is often instructed what kind of events to record and when or how often to do so. The diary method was adapted for use in HCI nearly two decades ago. One of the first scientific papers in HCI was the diary study presented by Riemann 1993 at InterCHI (Rieman, 1993). Since then, the method has received rather limited recognition by the research community in general. While there are obvious drawbacks of the method, such as the increased burden on the participant, we ascribe this lack of interest to the overall low application rate of longitudinal research in HCI. However, in recent years, there seems to be increased interest (as is the case for longitudinal research in general). This is also reflected in terms of recognition of the method in text books. “Research Methods in Human-Computer Interaction” by Lazar et al. (Lazar, Feng, & Hochheiser, 2009) is the first HCI text book to our knowledge that devotes a complete chapter to the diary method. The design and implementation of electronic diaries has furthermore made both the diary method and ESM much more flexible and accessible for researchers in various fields. These devices have added benefits, such as simplified data analysis (due to its digital nature) and the integration of richer data-gathering modalities,

such as audio and photo by means of electronic devices. In chapter 3.4 we present the PocketBee diary/ESM tool, which illustrates, among others, these issues in detail.

This chapter offers the following contributions to the field:

- First, we present an overview of the diary method. We will describe the origins of the method and discuss how it is applied in other fields, including the social sciences and psychology. We will illustrate the type of research questions that can be addressed with diaries by the means of the taxonomy presented in Chapter 2. Furthermore, we define a classification scheme of research designs that unifies diary and ESM research under one umbrella. In addition, we provide an overview of research studies in HCI that have adapted the diary or ESM approach, including one of our own studies, which we accordingly present in greater detail.
- We describe PocketBee, a multi-modal diary/ESM tool for longitudinal field research based on the Android mobile platform that allows researchers to remotely configure and react to data-gathering. The tool contributes to the field of electronic-diary tools by integrating an event architecture capable of supporting both diary and ESM studies individually or in combination. Additionally, a flexible and intuitive user interface allows the participant to gather data in multiple ways, e.g., through the means of questionnaires, text, voice-recordings, or photographs. The usability of the tool has been demonstrated in two case studies that will be presented. As an outlook for future work, we present a user interface concept for the researcher that allows the easy and ad hoc manipulation of underlying event configurations.

3.1 The Diary Method

As with most research methods in HCI, the diary method has been carried over from other research disciplines, such as psychology and the social sciences. Currently, however, we are not aware of a comprehensive work concerning the diary method in HCI in terms of the underlying thought model, the research

questions that can be addressed, the various possible research designs, or the data analysis that should be considered. In order to bridge this gap and provide a more extensive view, we will first discuss the work on the diary method that has been carried out in psychology and the social sciences. It is interesting to note that until a few years ago, these domains also lacked a comprehensive overview. Bolger et al. (Bolger, Davis, & Rafaeli, 2003) is one important such work in the domain of psychology research and Alaszewski (Alaszewski, 2006) in social research. The latter is described as the first comprehensive book on the diary method in the social sciences.

Alaszewski defines a diary “as a document created by an individual who has maintained a regular, personal and contemporaneous record” and provides four main criteria for a diary (Alaszewski, 2006):

- Regularity: diary entries should either be made at fixed time intervals (e.g., each day) or linked to specific events.
- Personal: a diary consists of entries made by one specific and identifiable person. It may be accessible by others who may read it but should not be allowed to contribute to it.
- Contemporaneous: The entries are made as soon as possible after the events or activities they report on.
- A record: A diary entry records what the participants themselves (or through the guidance of a researcher) find relevant and important. These can include events, activities, interactions, impressions, and feelings. As time is an important aspect for diaries, time often serves as a structuring tool for the record, which may be text but could also include audio or audiovisual data.

Alaszewski also distinguishes between unsolicited diaries and solicited diaries. Unsolicited diaries are those that are not imposed by researchers but are *real* diaries, not originally meant for research. These are especially interesting for historical researchers and biographers. Solicited diaries, on the other hand, are those that are imposed by researchers on the diarists, who are therefore participants in a study. With regard to the diary method in HCI, we are most interest-

ed in solicited diaries. Therefore, when speaking of diaries in the following chapters, we are referring to solicited diaries.

3.1.1 A Brief Introduction to the History of Diaries

The diary method is somewhat different from other research methods in terms of its history and theoretical foundation. Diaries were not invented by researchers, obviously, but are a cultural tradition and artifact which dates backs at least to 1000 BC. Alaszewski (Alaszewski, 2006) gives a short overview of the history of diaries as a (most often) personal notebook. He cites Sei Shonagon's *Pillow Boo*k and Murasaki Shikibu's diary as representatives of diary literature from the tenth century that are still existent, at least in the form of copied versions. In case of Shonagon, Alaszewski reports that the Emperor gave this diarist paper as a gift in the year 994 and she began to record "odd facts, stories from the past and all sorts of other things" (Morris 1970, p11, cited in (Alaszewski, 2006)). An important difference to official documents is the subjectivity, personal style, and interpretations that characterize these diaries, also apparent in the following quote from Shonagon:

> When the Emperor returned from his visit to Yawata, he halted his palanquin before reaching the Empress Dowager's gallery and sent a messenger to pay his respects. What could be more magnificent than to see so august a personage as His Majesty seated there in all his glory and honouring his mother in this way? At the sight tears came to my eyes and streamed down my face, ruining my make-up. How ugly I must have looked.
>
> (Morris 1970, p11, cited in *(Alaszewski, 2006, p. 3)*)

In Europe, diaries started to appear as personal records during the sixteenth century, such as that of King Edward VI. By the seventeenth century, diaries had become more widespread, as reading and writing skills among the population developed in parallel to the distribution of paper bound in books, ready to be written on. It was mainly the upper class in addition to monasteries that possessed both these skills and the materials. The personal motivations for keep-

ing such personal records were manifold. In the case of King Edward, for example, it was meant as an educational and formal exercise for his tutors. Similarly, scientists kept records of experiments and observations. In both cases, personal notes sneaked in from time to time or even predominated. One reason behind this is that writing a personal diary allows a level of reflection that is difficult to achieve without some form of externalization, and a diary is just one structured form for this purpose. Furthermore, in some social classes it became common practice to keep diaries. In later years, famous people in particular wrote diaries as a form of autobiography, always with the idea in mind of publishing these memoirs at a later time. Today, diary-keeping in a structured, day-to-day fashion has become quite rare, at least on paper. Instead, it seems that this kind of information is stored in blogs or on Facebook[4], reducing the level of privacy but also combining aspects of personal notes with the possibility of publication of one's thoughts.

Historical researchers in particular (including contemporary researchers and biographers) have begun to value the information contained in diaries as a source of additional perspectives. The biographer Ben Pimlott wrote an article for the newspaper *The Guardian* about the value of the information in diaries; he concludes:

> Diaries tell the truth, the partial truth, and a lot more beside the truth. Describing the same events. In them, you seek - and often find - an atmosphere, a sense of the mood of the moment, which cannot be acquired in any other way. They should never, ever, be taken as the last word. But as raw material for the reconstruction of the past, they are as invaluable as they are savagely entertaining.
>
> (http://www.guardian.co.uk/books/2002/oct/18/redbox.politics)[5]

[4] http://www.facebook.com

[5] Online available on 30.07.2011

Researchers find those documents invaluable for understanding people and their contemporary settings, as they provide insights that are normally not kept in official documents: insights about normal people and their daily lives, little details, or as Bolger et al. describe it, they capture “life as it is lived.” Even as researchers acknowledged the tremendous value of diaries, however, they strove to be able to gather this data purposefully and with control; thus, the diary method was born. In this way, the theoretical foundation of diaries – their ability to capture the events of our daily lives, providing insight into social and also psychological processes – is grounded directly in empirical data from the real world.

3.1.2 Research Questions and Types of Diaries

“Diaries are excellent for studying temporal dynamics” (Bolger, Davis, & Rafaeli, 2003)

As already stated, diaries are an inherently longitudinal research method and therefore they are especially suited to study change processes over prolonged periods. The beauty of the diary method is its flexibility to gather both rich qualitative data and quantitative data by means of rating scales. With regard to the research questions presented in Chapter 2, diaries can be the data-gathering method of choice in each case. Using diaries, one can easily aggregate data across time as well as specifically examine pre-post comparisons or the outcomes of change processes. As the data-gathering schedule can easily be specified, interest in the process of change or event occurrence can also be satisfied. In addition, the possibility to gather rich in-depth data in its spontaneous context allows insight into the Why and How of change processes. However, one has to be aware that diaries in principal are an indirect data-gathering method which relies on reported data and does not allow direct measurements. As we will show in section 3.4, electronic diaries stretch these boundaries through the possible integration of sensor data.

3.1.2.1 Quantitative Diaries

From a psychological and therefore mostly quantitative perspective, Bolger et al. (Bolger, Davis, & Rafaeli, 2003) identify the aggregation over time (referring to our “interest in the average over time”) and the modeling of time (referring to “interest in the shape of change” and specifically multi-level growth curve modeling as data analysis technique) as the two foremost research questions that can be addressed. Diaries can easily be equipped with standardized and precise rating scale questionnaires that allow the gathering of quantitative data over time. Thereby, diaries provide a test instrument for experimental and survey research that (1) can be integrated in the natural environment, (2) is less intrusive than direct observation, and (3) allows assessment of subjective feelings with reduced retrospective bias. The great advantage of diaries for modeling time is their high fidelity and flexibility compared to standard data-gathering methods like surveys. It is in principle possible to ask participants complete a diary entry at time intervals ranging from every two hours to once a month, allowing the researcher to investigate temporal dynamics at a very fine granularity (e.g., whether someone feels different in the morning than in the evening). For example, we could ask whether participants rated their mobile phone in terms of attractiveness differently when they bought it compared to four weeks later, and whether the subjective attractiveness differed on weekdays compared to weekends (when the subjects may have more free time to play with it or show it to friends). The key advantage of a diary here is the fine granularity of the data that can be obtained. First, this allows researchers to achieve a higher level of confidence in the data, as they can control for anomalies in the data, such as random peaks or lows. In addition, this technique does not miss dynamics in between the start and end points of the measurement, which allows for more complex research questions. Thus, for example, we can analyze what a typical rate of change is and its shape (e.g., whether it is at all linear) and also how people differ in their rate of change.

3.1.2.2 Qualitative Diaries

Alaszewski (Alaszewski, 2006) specifically points out the value of diary research for naturalistic and explorative research, investigating the Why part of our “Why

and How" research question. "Naturalistic" means that the researcher attempts to study the world as naturally as possible, by becoming part of the natural setting. Artificial settings such as experiments or even formal interviews are avoided. Thus, diaries are obviously well-suited, as no formal intervention is necessary once the study has started. Bolger et al. (Bolger, Davis, & Rafaeli, 2003) state that diaries facilitate "the examination of reported events and experiences in their natural, spontaneous context," which makes them "the document of life par excellence, chronicling as it does immediately contemporaneous flow of public and private events that are significant to the diarist" (Plummer 2001, as cited in (Alaszewski, 2006, p. 48)). As the focus is on qualitative data and understanding people, such diaries are much less structured and normally avoid rating scales or the like in favor of broader questions or different means of data-gathering than text. Participants are asked to report in their own words, e.g., how they would feel about their intimacy today or how they would describe their experience with their mobile phones. The result is highly individual diaries, most notably involving small stories or *bits of life*. These possess the ability, as Alaszewski notes, to provide insight into "taken-for-granted activities" or "tacit knowledge" – aspects of their lives that people have difficulty remembering or articulating in an interview situation, as they do not easily leap to mind. In addition, such a diary often includes not only descriptions of events or feelings, but also the participant's interpretation and reasoning without any interference on the part of the researcher. This further allows an exploration and understanding of the *Why* that cannot easily be derived from other research methods.

3.1.2.3 Interview-Diary

Alaszewski cites Zimmerman and Wieder 1977 (Alaszewski, 2006, p. 77) as the first to systematically combine the use of diaries with interviews. The basic idea is to establish a triangulation of the diary method with interviews. This seems quite straightforward, especially in naturalistic research in which a more open structure is used for the diary; this means that participants have quite a lot of freedom to express their observations or thoughts in their diaries and are not restricted to rating scales or closed-ended questions. Such data tends to be difficult to analyze, as aspects might be incomplete or might raise new ques-

tions. In the approach by Zimmermann & Wieder, the researchers therefore asked the participants to come to a debriefing interview some time after they returned the diaries. In the meantime, the researchers analyzed the diary entries and used them as the basis for the debriefing interview. The participants were then asked to expand on the diary entries or were asked for more background information about meanings or entries that were difficult to understand from the researchers' perspective. The requirement for a useful combination of interviews and diaries is that the researcher has time to analyze the data before the interviews take place. If this is feasible, this combination can provide tremendous added value, also in the context of HCI: for example, when participants are asked to report usability issues in a diary. The debriefing interview allows elaboration on these issues and also puts them into perspective; while a diary entry might be clearly motivated by emotion, the debriefing interview allows the gathering of a more reflective perspective on the subject. However, the task of balancing these possibly contradicting views is not an easy one.

3.1.2.4 Elicitation and Feedback Diaries

In HCI, Lazar et al. (Lazar, Feng, & Hochheiser, 2009) have adapted the interview-diary with their concept of elicitation diaries. They define these as diaries which specifically ask participants to capture events and situations they find interesting as a means of memory aid for an upcoming interview that concludes the study. An elicitation diary is therefore not a simple combination of the diary method and interviews, but rather a diary method that serves the (singular or foremost) purpose of providing memory props for the interview. If possible, such a diary approach should make use of multiple modalities, as in (Carter & Mankoff, 2005), which enriches the data-gathering process and allows participants to assemble visual props that help them remember situations. Lazar et al. contrast this approach to *feedback diaries*. Instead of the participant gathering memory aids or aspects they find interesting, the researcher should give much clearer guidance on what to report and how, e.g., by providing a structured template for data-gathering. This allows a much easier quantification of the data, which in return provides means of assessing research questions involved with the shape of change or pre-post comparisons. However, we find both the

notion of elicitation diaries and feedback diaries to be somewhat artificial; Lazar et al. also state that in most cases a hybrid form of the two should be applied.

3.1.2.4 The Experience Sampling Method

The Experience Sampling Method is an approach closely related to diaries. As we will show in the discussion of diary designs, it fits in nicely into the "Diary Universe." As the method has been one of the most often applied diary approaches in the social sciences, we will discuss it in more detail. The basic goal of ESM is to allow researchers in the social sciences to study "the quality of people's everyday lives – of what they do and how they feel about it" (Hektner, Schmidt, & Czikszentmihalyi, 2007, p. 3). Hektner et al. describe the approach as based on the idea of systematic phenomenology, which was primarily developed by Mihaly Csikszentmihalyi, one of the founders of ESM. His goal was to build a bridge between phenomenology, which is concerned with how people perceive things and how they are represented in consciousness, and traditional behaviorism, which in contrast looks solely at the outcomes of these mental processes in human actions. Systematic phenomenology therefore extends basic phenomenology by examination of how mental processes are expressed by means of empirical methods. ESM is the method to achieve this goal. The method was first developed at the University of Chicago in the early 1970s to study the experience of flow (Hektner, Schmidt, & Czikszentmihalyi, 2007, p. 7), in an attempt to resolve some of the issues the authors encountered with traditional daily diary reports. These reports were generally boring daily summaries, but Csikszentmihalyi was much more interested in fresh experiences and "stream of consciousness," as Hektner et al. call it. The basic idea to resolve the issue was to equip participants with a signaling device, such as a pager. At randomly signaled intervals several times a day during the week, participants would be asked to report on their current emotional state, feelings, or their physical or social context (for a more complete list, see (Hektner, Schmidt, & Czikszentmihalyi, 2007, p. 8)) by means of closed- or open-ended questions (although closed-ended questions seem to be the preferred option). Participants were to respond immediately upon the signaling event, thereby minimizing any

retrospective bias and capturing the current situation instantaneously. Hektner et al. describe the major benefit of the method as follows:

> [The] unique advantage of ESM is its ability to capture daily life as it is directly perceived from one moment to the next, affording an opportunity to examine fluctuations in the stream of consciousness and the links between the external context and the contents of the mind.
>
> *(Hektner, Schmidt, & Czikszentmihalyi, 2007)*

Hektner et al. discuss, that this technique is best used when participants are asked to report on everything in their daily experiences; thereby, the researcher gathers a completely random selection of these reports. However, in reality, the method has been often used for much more focused research questions; this narrows any perceived difference from the diary method in general. From the perspective of Hektner et al., there still are some significant differences to diaries: the latter often focus on activities and time use and not so much on immediate experiences. As Hektner et al. view diaries in the sense of filling out a daily report, they also argue that ESM reduces retrospective bias. As we will see in the following chapters, today's diaries are much more often integrated with mobile electronic devices and facilitate on-the-spot data-gathering quite similarly to ESM. However, ESM puts specific effort into instructing participants to respond promptly and to not postpone the experience reporting. Therefore, the approach can be considered to be stricter with respect to this specific issue.

The basic limitations of the approach are similar those of the diary method in general, which we will discuss at the end of this chapter in detail. Hektner et al. specifically mention the increased burden on participants and resulting problems of selective non-response and self-selection bias (Hektner, Schmidt, & Czikszentmihalyi, 2007, p. 7). The strict rule to react immediately upon signaling events further increases the burden in comparison to more relaxed diary studies. As the method relies on technical equipment to signal participants, the costs of implementation are also relatively high when a large number of participants take part in a study.

Overall, we feel that ESM is a tremendously useful approach to diary research and as such it is probably the most successful implementation of a diary method. In addition, often there are exceptions to the strict rules (e.g., postponing signals in inconvenient situations), or the approach has been used in combination with more traditional diary data-gathering. As we will show in the next section, from a research design perspective, the ESM approach fits very well into a holistic diary research paradigm.

3.1.3 Diary Research Designs

Diary designs, in terms of longitudinal research designs, are inherently prospective longitudinal panel designs. However, there are still a variety of possibilities of how to implement such a design in the case of the diary method, specifically regarding the data-gathering waves, as they are the sole responsibility of the participants. We will first present two of the most commonly referred to designs: that of Wheeler and Reis (Wheeler & Reis, 1991) and that of Bolger et al. (Bolger, Davis, & Rafaeli, 2003) (basically an extension of the former). Eventually, we will present a further extension to this scheme, unifying diary and ESM research and also taking into account the requirements of HCI research and the capabilities of modern electronic devices.

Wheeler and Reis (Wheeler & Reis, 1991) introduced a threefold classification, distinguishing between interval-contingent recordings, signal-contingent recordings, and event-contingent recordings. In the following section, we present the extension of this classification as discussed by Bolger et al. (Bolger, Davis, & Rafaeli, 2003), who distinguish between time-based and event-based designs.

3.1.3.1 Time-Based Designs

Interval and signal recordings describe designs in which time is used to define when a diary entry has to be made. The difference is that in the latter, the participant is not necessarily aware of the intervals between entries, or the intervals may be random. This is achieved by using an (additional) signaling device that notifies the participant of when to do a diary entry. Bolger et al. combined these

two designs to form time-based designs that may either use fixed or random intervals or a combination of both. A signaling device, while mandatory for random intervals, can also help the participant remember to complete a diary entry at fixed intervals. Typically, if it is important to catch the spontaneous reaction of a user, a random-interval design is the sensible choice. Such a design asks the participant to complete a diary entry at random, not at predictable time intervals, as indicated by a signaling technique. This is especially suited to study "internal" phenomena such as momentary experiences and psychological states (as in ESM). Bolger et al. use the study of the "frequency of stress experiences among students approaching an exam" as an example here. A researcher would try to avoid participants "preparing" for the diary reports or thinking too much about their momentary stress level in anticipation of doing the diary entry – both aspects that can be observed in fixed-interval designs. What one wants to gather is an unbiased and truly instantaneous report of stress; a random schedule can provide this. However, the burden on the participant is in general increased as the diaries become more intrusive, more demanding, and much more difficult to integrate into one's daily schedule. This drawback can be counteracted by allowing participants to postpone an entry, although this would increase the recall bias again and may reduce the benefits of the random-interval design.

Fixed-interval designs ask participants to complete a diary entry at certain predefined intervals: every three hours, every day, or one a week, for example. This design is often preferable when the studied phenomenon is expected to happen on a regular basis that can be incorporated into the schedule. In addition, it provides benefits in data analysis, as it allows the modeling of time as a factor without the worry of the possible effects of unbalanced intervals on participant responses. It is also easier for the participant to incorporate such a diary into his or her daily schedule. For the intimacy example discussed earlier, a fixed-schedule design might be appropriate, because one would expect such a study to be interested in an ongoing experience rather than instantaneous feelings. Such a design might ask the participant to report on the intimacy level at 6:00pm each day, for example.

What is important to consider in fixed-interval designs is the size of the interval. A long interval, while being less of a burden for participants, can introduce a recall bias, making the diary less effective and less valid. Furthermore, such a design might simply miss out on certain events or processes. However, very short intervals can also cause problems. In addition to the increased burden on participants, the increased signal-to-noise ratio can be difficult to analyze. This basically means that researchers may miss slower-acting processes if the intervals are much shorter than the change within the phenomenon that the diary is supposed to capture (Bolger, Davis, & Rafaeli, 2003). In the smart-phone example, one would assume that the perceived attractiveness would not change every three hours, probably not even every day. Therefore, an overly short interval may mask long-term changes from the researcher. Furthermore, there is a danger of deadening regarding the participants' reactions towards the diary. If they are asked to answer the same questions in short cycles without including their personal impression of whether anything has changed, they might lose their motivation and dedication to participate in the study, as their actions would seem meaningless. Still, Bolger et al. recommend the use of shorter intervals instead of overly long intervals, as short intervals also allow the analysis of data with different time lags in order to identify changes.

In general, time-based designs are the preferred option for quantitative diaries and research questions that address the shape of change or pre-post comparisons, as they are better suited for modeling time and incorporating quantitative analysis, as stated above. Event-based designs, on the other hand, are preferred in naturalistic and ethnographic research, because they make it easier to understand the Why and How of change processes.

3.1.3.2 Event-Based Designs

Event-based designs, such as the event-contingent recording by Wheeler and Reis (Wheeler & Reis, 1991), do not use time to trigger diary entries but rather other external events. The basic idea is that whenever such an event occurs, the participant should complete a diary entry. Bolger et al. stress that the relevant events must be clearly pre-defined prior to the study, so that participants are well aware of when they should record a diary entry. This design is often

used for rare or isolated events, which are difficult to capture in time-based designs or through other research methods. The goal is to reduce the recall bias for such events as much as possible; longitudinal aspects are not of focal interest here. It is important that the triggering events are absolutely clear to the participant, as they may otherwise record irrelevant data (which might be difficult to identify as such). Bolger et al. suggest focusing on a single type of event instead of using multiple triggering events, in order to reduce any chance for ambiguity. There is still a risk that participants may fail to correctly identify an event. Researchers also risk overgeneralizing from the reported events: if the pre-defined events do not fully cover the phenomenon under investigation, than simply relying on reports about these events can lead to inconclusive or incorrect deductions.

3.1.4 A Unifying Classification Scheme for Diary and ESM Research

We agree with the above classification on a general level, and would especially like to stress that Bolger et al. do not conceptually distinguish between diary and ESM research. Even Hektner et al. (Hektner, Schmidt, & Czikszentmihalyi, 2007), while approaching this topic from the ESM direction, cite the diary designs by Wheeler and Reis (Wheeler & Reis, 1991) as one way to distinguish between different ESM designs. We lean more towards the view of Bolger et al. in terms of viewing ESM as a specific type of diary research: ESM is more specific in terms of research design, while the diary method allows much more flexibility here. However, even this is changing now, as more and more ESM studies (like diary studies) begin to make use of sensors other than time to trigger participants' responses. Since these kinds of designs are reflected neither in the classification by Bolger et al. nor by Wheeler and Reis, we will present an extended classification scheme that focuses especially on HCI and takes the technical capabilities of current electronic diary and ESM tools into account. Our classification scheme distinguishes more clearly between designs that are based on certain automatically triggered conditions and those that require the participants' judgment. We call the first type *condition-based designs* and the second type *human recognition-based designs*. In the following sections, we will

describe how these further diverge; stressing that for most studies, a combination of different aspects could be the most sensible approach.

3.1.4.1 Condition-Based Designs

These designs require the researcher to define certain conditions which then trigger the data gathering or response process. In all cases, data-gathering can be either manual (by the participant) or automatic (e.g., by starting a logging function on an electronic diary). We can further distinguish between time-based and sensor-based conditions.

- Time-based conditions include the classic ESM design with random signals asking the participant to enter responses regarding, for example, their current emotional state (completed in a paper notebook in “the early days”). These designs also include what Bolger et al. call fixed-schedule designs in which, for example, questionnaires are presented each day at 6:00pm.
- Sensor-based conditions extend these designs by including any kind of sensor data that can be used to define conditional events. For example, a GPS sensor could be used to create a condition that a participant should respond any time they are at a certain place. A CAN-bus sensor in a car could be used to ask the participants “what happened” whenever a “hard braking” is detected. Such conditions, as stated above, could also be used to trigger automatic data-gathering, perhaps by taking a screenshot of the mobile application or a photograph using a camera within the device. Furthermore, as modern mobile devices provide a variety of sensors, these conditions can be combined to create more complex conditional events.

The technological possibilities of mobile devices also allow the differentiation between what might be called modal/immediate dialogs and user-controllable dialogs. The former present the data-gathering or response dialog directly on the screen as soon as the conditions are met, taking precedence over every other application on the device. The advantage here is that the researcher can increase the probability that the participant will respond immediately. However, the researcher must also make sure that this does not happen in inappropriate situations, e.g., during a call. User-controllable dialogs notify the participant if

conditions are met, but might otherwise leave the device in its current state. The researcher therefore has to design a UI element that allows the participant to start the data-gathering/response dialog at a later time.

3.1.4.2 Human Recognition-Based Designs

In general, these designs also require the researcher to pre-define a “condition.” However, this condition is not definable via sensor data, thus the participants have to function as “condition-checker” from a technical perspective. This category includes the traditional diary design in which participants are asked to record data whenever certain situations occur, or possibly more or less complete logs of their daily life (e.g., (Czerwinski, Horvitz, & Wilhite, 2004)). The degree to which these “real-life” conditions are defined by the researcher can vary, as it depends on the research question of how focused one would want the participants on the data-gathering. In situations of explorative research, it might be beneficial to not overly restrict the participants, applying instead more of a participant-defined events design. However, it is still important that participants get a good idea of what is interesting to the researcher. Otherwise, the burden of deciding whether or not to record a situation might be too high and result in some things being recorded and others, although relevant, being missed. With regard to the user interface, the researcher must design a UI element that allows the participant to start the data-gathering/response dialog at any time.

In practice, it is very useful to combine condition-based and human recognition-based designs, not only within one study but also within one data-gathering situation. The input from human recognition-based designs could be used as “sensor”-conditions and trigger further response prompts. For example, whenever a participant takes a picture in a “human-recognition” situation and the GPS sensor-data shows that he or she is at home, the researcher could trigger a further response prompt asking the participant to make an audio message describing the image. On the other hand, it might be useful to restrict or control human recognition-based responses by combining these with sensor- or time-conditions (e.g., allowing a participant to record data only between 8:00pm and 10:00pm).

3.1.6 Advantages and Challenges in Diary Research

While we have already addressed some of the advantages and challenges of diary research in general, we would like to specifically summarize these in this section.

3.1.6.1 Advantages

One of the principal advantages of the diary technique as a research method is that it is capable of capturing data in its **natural and spontaneous environment** (Bolger, Davis, & Rafaeli, 2003). It can therefore be regarded as **unobtrusive** (Hektner, Schmidt, & Czikszentmihalyi, 2007, p. 7) or at least much less intrusive than research methods that require a researcher to be present. Overall, this may lead to a **higher ecologic validity** of the data (Czerwinski, Horvitz, & Wilhite, 2004). The reduced instrumentation should also reduce the Hawthorne effect or **reactivity/reactance** (Bolger, Davis, & Rafaeli, 2003), a common bias introduced by studies in which participants behave differently because they know they are being observed and studied. However, as we will see with respect to the challenges, reactance can be a factor that must be considered.

It is a very **flexible** method that allows the capturing of **qualitative and quantitative** data. Qualitative data can be gathered in the form of text and more recently by means of additional modalities, such as photographs, voice recordings, drawings, or even video (Gerken, Dierdorf, Schmid, Sautner, & Reiterer, 2010b). Quantitative data can be captured by means of questionnaires and test instruments based on rating-scales. In addition, recent electronic diary tools allow the capture of context information **through sensors and logging** data (Froehlich, Chen, Consolvo, Harrison, & Landay, 2007). By means of this data, the diary can gather data on factors such as perceptions, thoughts, feelings, activities, behavior, and context information (Hektner, Schmidt, & Czikszentmihalyi, 2007), among other factors, again illustrating the flexibility of the approach. With regards to HCI, the method can preserve the **mobility of participants**, thereby making it especially well-suited for studying interactions with ubiquitous technology.

Another advantage, according to Alaszewski (Alaszewski, 2006), is that a diary allows the researcher to capture also the **little details** of everyday life, aspects that might be difficult to observe or may be forgotten in interviews. It can capture data at a highly detailed granularity, as participants are asked to report **events when they happen**. This is related to the **reduced retrospective bias** of the method. Even in "daily" diaries in which participants report on their day in the evening, there is much less potential for retrospective bias in comparison to retrospective interviews. Recent e-diaries and ESM in general reduce the divergence between event and reporting even more, making **data-gathering nearly instantaneous**.

The method is also **easy to communicate** to participants, although clear instructions are still important and necessary (Bolger, Davis, & Rafaeli, 2003). Overall, most people are familiar with the basic idea of keeping a diary, and many have experience with activities similar to diary-keeping through use of Facebook and Twitter.

From the perspective of this thesis, as discussed earlier, the method is especially appealing, as it is **inherently longitudinal**. However, the researcher must still put in thought as to how the participant should capture data so that also change processes can be analyzed. Nevertheless, the basic features of the method are already fleshed out in such a way that they not only support a longitudinal design but also enforce it. Thus, even if the method is not always used to analyze longitudinal change processes (as we will see in the next section that presents several example diary studies from HCI), it may lead to future longitudinal studies as researchers begin to grasp the full potential of the method.

3.1.6.2 Challenges and Drawbacks

As with every research method, the diary method also presents several challenges and drawbacks, not all of which can be easily addressed. The most obvious one is the increased **burden on the participant**. The method basically shifts the data-gathering responsibility from the researcher to the participant, thereby shifting the effort linked with this activity. This burden includes **physical as well as cognitive strain**. The physical burden is related to the need to rec-

ord data in the designated way and to carry the diary around, no matter whether it is pen and paper or an e-diary (although the latter offers the possibility of integration into the personal smart phone of the participant). The cognitive strain relates to the fact that the participants in many diary designs must remember to actually do the data gathering, must think about how to do it and what to capture, might feel stressed by the data-gathering itself, and could even feel guilty if they get the impression that they are unable to satisfy the requirements (in terms of the amount of data gathered, for example). All this can lead to serious issues, one of which is the **panel attrition** seen in longitudinal research in general. In the case of diary studies, it can be assumed to be even more severe and difficult to observe – a participant may not explicitly drop out but simply reduce his or her efforts in data-gathering, an effect that Hektner et al. call **selective non-responses** (Hektner, Schmidt, & Czikszentmihalyi, 2007). This is related to the issue of **uncertain compliance** (Bolger, Davis, & Rafaeli, 2003), as researchers often do not know whether participants actually paid attention to the instructions and followed them throughout the study. For ESM-like studies, this means that there is uncertainty as to whether participants report their feelings or emotions immediately when prompted or postpone their response without letting the researcher know. Reduced compliance often is not a deliberate act, but is simply due to ignorance of the importance of compliance or forgetfulness on the part of the participants. The use of e-diaries has helped to improve this situation a great deal, although not by directly increasing compliance, as this is still an open debate (see, for example, (Broderick & Stone, 2006)). E-diaries allow the researcher to know much more about the context in which the data-gathering takes place (or does not take place). This starts with the knowledge of the exact time when a diary entry was completed. It may also allow the researcher to check whether activities being reported actually happened (by means of interaction logs that are automatically captured, for example).

As mentioned above, there is also an on-going debate on the influence of **reactance** (Bolger, Davis, & Rafaeli, 2003). For example, Czwerwinski et al. state that “individual behaviors may be altered because they know they are being studied” (Czerwinski, Horvitz, & Wilhite, 2004). However, several authors, as reported by Bolger et al., “argue that diaries may lead to less reactivity than

other forms of data collection because of a **habituation** process."[6] By "habituation," they mean that the diary itself may become a familiar artifact in one's personal life, thereby reducing any possible reactance effect as time passes. However, only a limited amount of research has actually investigated this issue. Additionally, habituation might occur in longitudinal research in general; whether a diary reduces reactance disproportional in comparison to other methods has not yet been determined.

On the other hand, **habituation** can also have a negative effect on diary studies (Bolger, Davis, & Rafaeli, 2003). Imagine a study that asks participants to fill out the same questionnaire every day. At some point, participants may stop putting in the necessary effort and just try to fill it out as quickly as they can. They might also overlook possible changes in the questionnaire, simply because they expect it to always be the same. In addition, for research questions that ask participants to report certain events, habituation can be problematic. While they may be motivated to report every little detail at the beginning, this thorough reporting will most probably peak and then drop off. Especially if they are asked to report on the same events all the time, they might leave out important details at some point, simply because they have the feeling that they already reported the event in enough detail earlier. Again, this effect can only be lessened by thorough instructions so that participants know what is important to the researcher. In addition, a good relationship between researcher and participant can help to keep participants motivated and involved. To this end, it might be necessary to include further feedback loops and meetings, even though they might not be used to gather additional data. The duration of a diary study is also an important consideration. For example, Hektner et al. (Hektner, Schmidt, & Czikszentmihalyi, 2007) claim that one week seems to be the optimal duration for ESM studies; most studies that we present in the next section from HCI research also tend to limit their duration to a few weeks. However, this depends on the level of detail that is required from the participants' reports and the frequency of reports.

[6] Bolger et al. cite Litt et al. 1998 and Gleason et al. 2001 as some of the researchers who support this argument.

We conclude this section with a serious issue inherent to longitudinal studies as well: **Construct validity over time**, which basically means that one has to be sure that the test instrument actually measures the same construct as time passes. This is also true for qualitative data and relates to the issue of habituation mentioned above: An event or situation might be reported completely differently on Day 1 of the diary study in comparison to Day 5, although the event itself may have been very similar and was also experienced similarly. In addition to habituation, Bolger et al. note that participants may simply achieve a more complex understanding of the problem domain, thereby altering the way they report. Bolger et al. also discuss what they call **"gradual entrainment,"** in which participants may adapt the view of the researcher as they are exposed to it via the diary (e.g., through the questionnaire, instructions, or feedback loop). This can also lead to problems in qualitative studies when participants start to only look for events that the researcher has described as potentially interesting, missing others that were not mentioned but may nonetheless be of equal importance.

Overall, it is important that researchers conducting diary studies be aware of these challenges and drawbacks. While it may not be possible to reduce or avoid all of them in a given study, knowing about their potential effects should at least lead to a critical review of the data gathered with respect to these issues and eventually help during the interpretation phase.

3.2 The Diary Method in HCI

While we have discussed research questions and research designs in theory, we think that illustrating the application of the diary method with real world examples facilitates comprehension. The idea of this section is to give a broad overview of the range of research questions and specific study designs for which the diary method can be used in HCI research. Obviously, the selection is not meant to be comprehensive, but rather diversified and inspiring. In addition to citing the work of other researchers here, we will also present one of our own diary studies in detail in Section 3.3. It is important to note that we did not limit

the studies examined (including our own) to those involving electronic diaries but include a wide range of different “implementations” of diaries.

3.2.1 The Diary Study (Rieman, 1993)

The first study we would like to cite is one of the first diary studies in HCI – in fact, it is often cited as being the very first. Riemann presented his specific diary approach as the missing link between lab-based experiments and observations in the field, as “objective tools for workplace investigations” (Rieman, 1993). Riemann suggests a very specific and detailed method for conducting a diary study, using preprinted log forms that must be filled out by the participant each day and that break each day into 15- or 30-minute intervals. Participants are then asked to record all activities during working hours or during the time span relevant to the researcher. In his example study, he applies a mixture of a time-based design with fixed intervals and a human recognition-based design. The fixed-schedule time-based design is implemented in such a way that the diary form had to be completed each day. With regard to the human recognition-based design, Riemann additionally equipped participants with a further template: the Eureka reports. Within these, they were asked to document every learning event they encountered (the study investigated how participants learned to use a computer). In addition, he conducted daily debriefing interviews, which were meant to discuss the categories for the activities participants had provided in the diary form. After one week, the study concluded with an overall debriefing interview. The study had several important results. First, the Eureka reports could serve as the basis for experiments that would be designed specifically to test the documented learning events. Riemann also notes that the personal interaction of the researcher with the participants during the study is a key factor motivating participants. He finds that participants filled out the forms quite differently; for example, one participant completely lacked detail, writing simply “programming” for the entire day. People often tried to focus their reports on things they thought could be interesting for the researcher. For example, the participant who left out all detail said that he “didn’t do anything you were interested in” and even apologized for not learning anything new: “I should learn

something so I can fill out a Eureka report." Riemann concludes that the researcher must continuously reiterate what people should do or not do and thereby try to counteract such biases.

Riemann's study is a prototype paper for the diary method, as it provides a many insights into the practical application of the method. While we do not agree with some of the very strict and specific design decisions in his diary method, we understand that as it was one of the first to use the process in HCI, he had attempted to provide a replicable method.

3.2.2 A Diary Study of Task Switching and Interruptions (Czerwinski, Horvitz, & Wilhite, 2004)

The study by Czerwinski et al. focuses on characterizing how people twist multiple tasks when interrupted. The problem space here is that information workers must often switch between tasks, and in this process some tasks are forgotten. They hypothesized that task-switching was a principal reason behind this effect. Czerwinski et al. discuss several ways of how this could be studied, e.g., by videotape analysis, which had the obvious drawback of only being able to observe what happened without knowing why it happened. They conclude that a diary study could tremendously increase the ecological validity, although they were aware of problems such as the increased burden on participants and possible bias because of reactance. They implemented the diary by use of existing digital tools, primarily an Excel spreadsheet with one worksheet for each day and one for additional instructions. Each row in the spreadsheet represented one task; several columns were predefined in which participants had to fill in descriptions of the task, such as the start time or the difficulty of switching to the task. The participants were asked to fill out this Excel sheet during their normal activities and to document any task-switching. Thereby, the authors implemented a human recognition-based design. Unfortunately, the authors do not discuss any shortcomings of the method – for example, whether people managed to fill out the Excel sheet immediately after switching to a different task or not. In addition, one could argue that by asking the participants to switch to the diary task, the authors introduced an additional bias, as task-switching was itself the

focus of the study. However, the paper does utilize a different type of electronic diary that is especially useful for studying or evaluating desktop applications in HCI.

3.2.3 An online forum as a user diary for remote workplace evaluation of a work-integrated learning system (Lichtner, Kounkou, Dotan, Kooken, & Maiden, 2009)

Lichtner et al. use yet another way of incorporating a diary into pre-existing tools. For their analysis of a knowledge management platform, they use a phpBB forum as a diary, with each participant having an individual forum as well as a shared space to report on any problems they encountered with the platform. This again resembles a human recognition-based design. The authors manipulated the forum software in such a way that each new forum post was pre-filled with prompts and questions. However, participants were free to ignore these. They were also given the option of adding screenshots to their posts. The forum entries were then analyzed and categorized in detail, showing types such as usability issues and technical notifications, but also communication with the researchers or other participants. In the shared space, participants were allowed to help each other. In contrast to the first two diary studies presented in this chapter, this had a much more explorative character, with the goal of finding evidence for possible redesigns of the knowledge management tool. As a result, this study does not present a diary method used for a specific research question, but rather as a usability evaluation method.

3.2.4 Mobile taskflow in context: a screenshot study of smartphone usage (Karlson, et al., 2010)

Karlson et al. were interested in a research question similar to that of Czerwinski et al., as they also studied task interruptions. However, they did not restrict this to the workplace but were especially interested in task interruptions in mobile scenarios. They therefore made use of participants' mobile phones as diaries (iPhones and Windows mobile devices); these were equipped by the re-

searchers with a screenshot tool, incorporating a mixture of human recognition-based design and time-based design with fixed schedules and also implementing what Lazar calls elicitation diary (Lazar, Feng, & Hochheiser, 2009). Participants were asked to take a snapshot whenever they were interrupted during a task. This simply required pressing a button and no further interaction with the diary application. Every evening (time-based design) they were instructed to upload all screenshots taken, annotate them, and rate their frustration and urgency on a 5-point scale. The human recognition-based design thus allowed reduction of any memory bias; by minimizing the task to a simple button press, the researchers could also reduce any further “task-switching bias” the study itself may have caused. A debriefing interview at the end of the study involved the participants actively in the coding of the data, namely, in finding categories for the screenshots and activities. Karlson et al. further provide additional background information on the study details; for example, they relate that participants were instructed with the help of scenarios, which were kept quite broad so to not overly restrict participants. In addition, to increase motivation, they included a variety of incentives. A fixed amount of money was given out for the two interviews (one at the start of the study and the final interview). Participants were also rewarded monetarily for the number of diary entries they made and for being the top diary entry maker. This type of monetary motivation cue always poses the danger of introducing a bias, as participants might start making diary entries without any substance. Since they would probably not want to acknowledge this, they might come up with invented explanations for their entries that could mislead the researchers. One interesting aspect of this study is the seamless integration into participants’ daily lives by making use of their own mobile phones. This definitely helps to raise the acceptance bar for such studies among participants and to further reduce the perceived level of instrumentation.

3.2.5 “It's just easier with the phone" - a diary study of Internet access from cell phones (Nylander, Lundquist, Brännström, & Karlson, 2009)

Nylander et al. were interested in studying how people use the mobile Internet on their smart phones. Interestingly, they relied on pen and paper diaries despite their interest in mobile phone usage. The pre-prints included examples of what to record and several data fields that structured the diary, including information such as time stamp, duration, location, which web page or application used, and so forth. Unfortunately, no information is given about when participants were asked to record these diary entries and in what format the diary pages were (e.g., was it possible to carry them around and were people asked to do so?). One of their main findings was that people mostly used the Internet function at home. One could hypothesize that the pre-printed diary forms were easier or more often available when they used the Internet at home compared to using the Internet on the go. However, the authors do not comment on this issue. Another interesting aspect, in line with all the presented diary studies so far, is that the researchers did not specifically look for changes in the data over time but instead aggregated the data, focusing solely on the first research question within our taxonomy: interest in the average person. A follow-up to this study is the study by Hinze et al. on mobile information needs (Hinze, Chang, & Nichols, 2010), again incorporating a paper-based diary. This time, the authors explicitly describe the pocket-sized form of the diary.

3.2.6 Data Logging plus E-diary: towards an Online Evaluation Approach of Mobile Service Field Trial (Liu, Ying, & Wang, 2010)

One example of an electronic diary in combination with logging is provided by Lui et al., who studied the use of online mobile services. A logging implementation automatically logged every interaction of participants using a service on a mobile phone – a Nokia device that was handed out to each of the 82 users for 2 months. The diary itself was accessible through the web and asked participants to report on the location, features, and any feedback they might want to

give. Unfortunately, the study gives little detailed information about the specific research design or instructions. The authors also failed to analyze logging and diary data in combination, to investigate in the compliance of participants, for example. In the following section, we will show how diary and logging data can complement each other nicely and allow the investigation of such aspects.

3.2.7 Conclusions

The presented studies demonstrate that diaries in HCI have been applied quite differently, and that even in the year 2010 some researchers are still reliant on pen and paper techniques. A considerable problem is that many studies lack details in terms of the specific research design they use for the diary study. We hope that by means of our classification scheme, researchers will have an easier way to classify their study and thereby to help others understand the approach taken and learn from it. Another significant outcome of this section is that all the studies focused on research questions interested in the average person or aggregation over time (as does our own study presented in the next section). We assume, as is generally true for longitudinal research, that there might be several reasons for this. First, diary studies in HCI seem to be used much more often to gather rich qualitative data, making it more difficult to analyze changes. In psychology, these studies are mainly used with pre-defined survey test instruments that permit easy quantification and thereby allow the use of statistical analysis methods as presented in Chapter 2. Second, designing a study to specifically analyze change processes takes more preparation and greater knowledge about the subject of the study. However, most diary studies in HCI seem to be of explorative nature with little pre-knowledge about what data to expect and how the data could change over time. Third, diaries are one of the easiest ways to extend a field study over a prolonged time period in order to achieve higher ecologic validity. Therefore, many researchers in HCI look to the diary method for this very reason and not to specifically analyze change processes.

3.3 HyperGrid vs. HyperScatter: A Multi-Dimensional Longitudinal Case Study

Parts of this chapter were published in (Gerken, Demarmels, Dierdorf, & Reiterer, 2008b) *as well as in* (Gerken, et al., 2009b).

In this section, we present a study that aimed at evaluating the usability of two different information visualization tools for searching in a movie database. The diary method was one important component of the study; however, the study also showed the importance of triangulation – in this case, diaries with automatic interaction logging and interviews. We will briefly present the research question and the design of the information visualization tools to allow the reader to comprehend the results more easily. We will then focus on the design of the study and the triangulation of the three methods in particular before discussing the results, again focusing on the effect of the triangulation, especially on how diaries and log files can complement each other to account for non-compliance.

3.3.1 HyperGrid and HyperScatter – Visual Information-Seeking in a Movie Database

The challenge of information-seeking is as old as civilization itself. Even the very first document archives dating back to 2600-2400 BC (the so-called room L. 2769 in the ancient Syrian Ebla) organized tablets in chronological order or by genre. Together with an arrangement similar to file cards, this allowed a quick scanning of the documents (Trumble & Marshall, 2003). The flexibility of access was increased substantially with the introduction of indices, keywords, and metadata. These enabled librarians to directly access documents without being limited by the existing order in physical space. In 1876, in his “rules for a dictionary catalog” Cutter defined that a catalog should allow one to see what the library owns and to search a document by different metadata (e.g., author, title, subject), and should assist the selection of a book or document by adding bibliographical and literary records (Cutter, 1876). With the introduction of online public access catalogs (OPAC) in the 1980s, the burden of information-seeking was passed from trained search mediators or librarians to the end-users, who

started to directly interact with catalog systems. Finally, the arrival of the World Wide Web introduced entirely new concepts of "digital libraries," which – in accordance with Borgman's definition (Borgman, 2003) – no longer limited themselves to the safe ground of well-kept single information spaces. With the World Wide Web, information-seeking has quickly expanded into the "wild," now spanning public and personal information, local and remote devices, professional and user-generated data, and a vast amount of miscellaneous content. Today's information-seeking systems are therefore turning the traditional information spaces of the past into a "personal information cloud" of an as yet unknown quantity, dimensionality, and heterogeneity, constantly expanding, organized and arranged by the user (Kaptelinin & Czerwinski, 2007). In such a context, search is an essential task in the workflow. The HyperGrid and HyperScatter visualizations have been part of the MedioVis project, which sought to simplify the use of information-seeking systems for novice and casual users. Since 2004, it has been accessible on more than 150 workstations in the library of the University of Konstanz. It allows users an alternative approach to the standard online catalog system for searching through more than 70,000 multimedia objects such as movies or documentaries. One alternative scenario includes titles from the international movie database (imdb.com) instead. This basically provides a different GUI to the imdb.com website and allows users to inform themselves about movies and the people involved in movie-making (actors, directors, producers, etc.), with access to movie metadata such as ratings, reviews, and cast lists. This alternative scenario was used in this study, thereby enabling participants to use the software at home and to access much more information about movies in comparison to the limited library data.

3.3.1.1 HyperGrid

The basic idea of the HyperGrid is to merge the concepts of a table-based visualization with details on demand and browsing functionality, integrating a direct-manipulative zooming interaction (Jetter H.-C. , Gerken, König, Grün, & Reiterer, 2005). The HyperGrid groups all attributes of a document or object by three aspects of interest. The grouping is done based on semantic similarities between attributes and is modeled in our attribute space concept. At first

glance, the HyperGrid looks like a standard table with each aspect of interest represented in one column. Users are able to zoom into a cell, however, resulting in an enlargement (see Figure 22). Depending on the zooming duration, which directly corresponds to the users' degree of interest, more attributes appear seamlessly. This underlying technique of a semantic zoom was first introduced by (Perlin & Fox, 1993). While this is a very intuitive and direct interaction, it also allows us to integrate heterogeneous attributes such as images or video clips into one coherent visual presentation. Furthermore, we are able to merge an attribute or metadata-focused view (the table) with an object view. Users can zoom into a cell until they reach the object itself, which is presented in an overlay window. This window only covers the cell, thus the context is still preserved. External information spaces can also be integrated (e.g., standard web sites, but also web services such as GoogleMaps). To allow dynamic filtering, we integrated a table-filter concept that allows simple keyword filtering in the headings of each column. For easier comparison of different attributes, a user-adjustable additional column allows users to grab attributes that are deeply hidden in the information space.

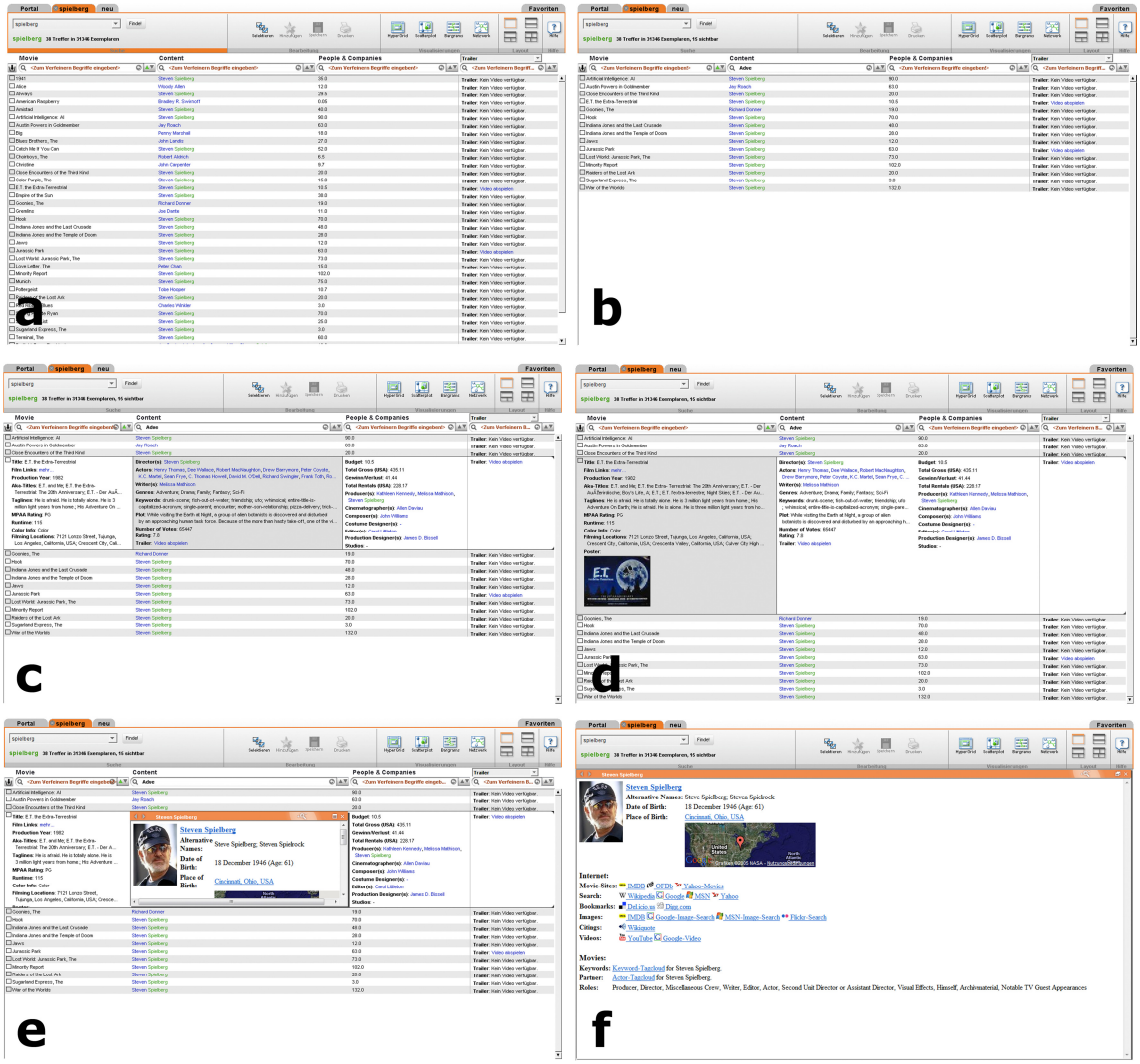


Figure 22: The HyperGrid visualization. Images a) to f) show different zoom levels as one zooms into an individual cell of the grid.

3.3.1.2 The HyperScatter

The HyperScatter was developed as an alternative visualization. The HyperScatter is a zoomable two-dimensional scatter plot that allows an overview and the exploration of correlations between quantitative or categorical data (see Figure 23). It further supports the effective selection, zooming, and filtering of user-defined subsections of the plot and therefore especially supports quantitative filtering and reasoning. To access details on demand, it integrates the same interaction concept as the HyperGrid. Clicking on a data point triggers an animation; the corresponding detail information is displayed along the different aspects of interest in an overlay window. We think that such a zoomable scatter

plot is especially suitable when screen space is limited (for example, on mobile devices such as smart phones or PDAs).

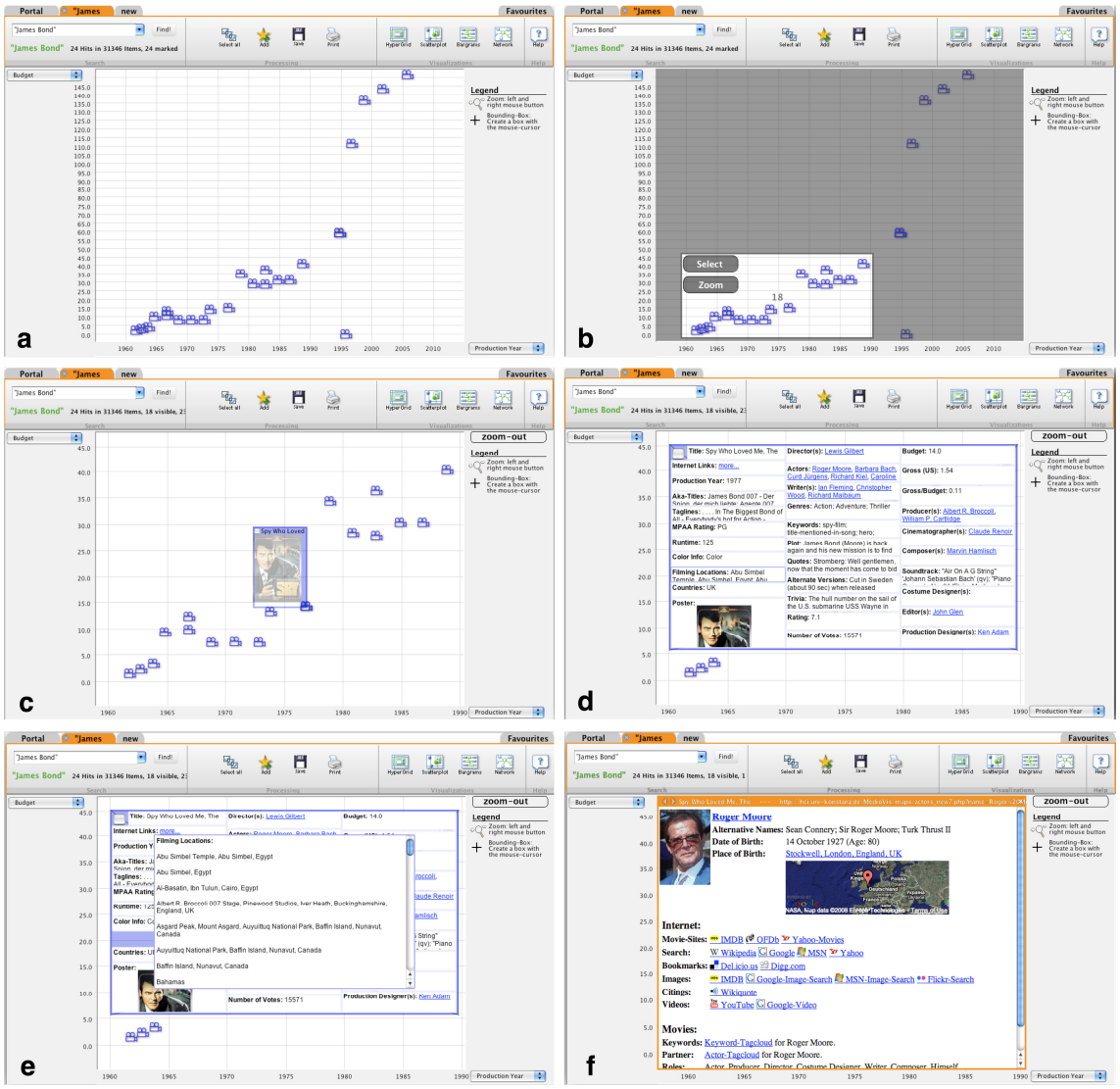


Figure 23: The HyperScatter visualization. Images a) to f) show different zoom levels after zooming into one specific object in the scatterplot.

3.3.2 Research Questions and Study Design

The goals of the study were a comparison of the HyperGrid and HyperScatter visual seeking tools in terms of their usability and their practical usefulness in a real-world scenario, not just in the controlled environment of a usability lab. We were also interested in whether the HyperScatter concept could serve as a stand-alone visualization, or whether it should be combined with other tools. In

accordance with the focus of this thesis on longitudinal research, we chose a longitudinal study design that allowed participants to learn how to use the visualizations and to use them in a real-world setting. With regard to the taxonomy presented in Chapter 2, we were interested in the average usability issues over time and did not focus on the analysis of changes. In the study, we implemented both techniques as stand-alone systems with access to a part-mirroring of the Internet Movie Database, including several thousand movies as the data basis.

3.3.2.1 Design and Participants

For the study, we selected eight participants from the University of Konstanz, students of varying majors. The participants were selected based on their self-stated high interest in movies and cinema in general. Thereby, we were able to assume a higher level of intrinsic motivation to use the tools. The study lasted a total of two weeks: one week for each of the two tools. Our participants were therefore able to work with both systems, compare them, and judge them accordingly. As stated above, we used a triangulation of interaction logs, interviews, and diaries for data gathering. We will discuss these in more detail below. The interviews framed our study with one start-up and two debriefing interviews (one after each week, see Figure 24). During the start-up and the first debriefing interview, the next week's tool was installed on the participants' laptops. They were allowed and encouraged to contact the researcher whenever they needed any kind of assistance and were given the researcher's email address and phone number. After about three days of the free usage of a tool, participants were handed a "weekly-task" The idea of this task was to motivate users to use the tool and also to guarantee comparable usage time for all participants, including the key functionality. The task asked participants several movie-related questions; to answer the questions, they were supposed to use the HyperGrid or HyperScatter. We also alternated the order in which participants used the two tools: half of them started with the HyperGrid and the other half with the HyperScatter. However, in the second week, all participants were allowed to use both systems. This was done to allow participants to directly compare the systems and also state an implicit preference for one of the tools.

However, the “weekly-task” still had to be done with the week-specific tool that had not been used in the first week. Participants were rewarded with 20 EUR at the end of the two-week study.

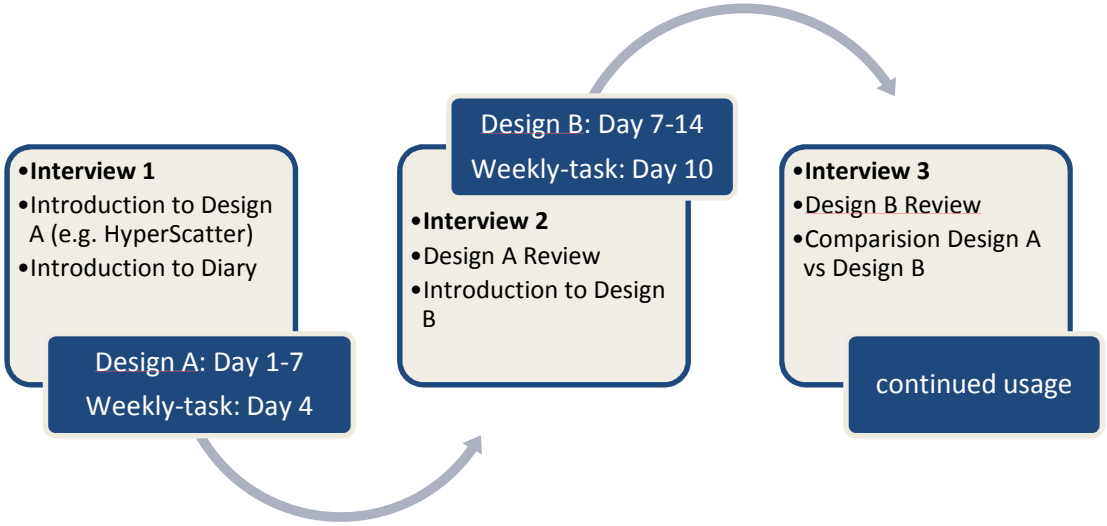


Figure 24: Study Design

3.3.2.2 Data-Gathering Methods

For the interaction logs, a logging technique was implemented that basically logged all interactions with the tools and directly transferred these to a server. There they were stored with the according time-stamp and user tag so that we were able to match all logs to our individual participants. The interviews were semi-structured. The first interview was used to introduce participants to the study goal and to introduce them to the first tool. It was also meant to create a bond between researcher and participant, in order to increase the motivation to participate in the study. During the second and third interviews, participants were asked to report any usability issues they had come across during use and to directly show them in the system. They also handed in the diary logs, which were scanned by the researcher during the interview to see whether any important issues had been left out by the participants. In the second interview, participants were also instructed how to use the second tool. While these interviews captured the reflected impression of usability, we used diaries to capture the *in situ* and spontaneous reactions and problems. We implemented a pen

and paper diary in an event-based design. The diary sheets were relatively highly structured (see Table 6). General fields for date, time, and ID allowed us to compare the diaries with the interaction logs. In addition, the diary was meant to help us find out the purpose of the tool use and uncover any usability issues. One aspect we were interested in from an interaction design perspective was the use of the external sources that are embedded within both the HyperGrid and the HyperScatter. We also had an additional 5-point rating scale for the level of fun people had while using the system and additional pros and cons. The design of the diary was borrowed from the Rochester Interaction Record by Reis and Wheeler (Reis & Wheeler, 1991). Participants were asked to fill in diary sheets on two occasions: first, whenever they came across a usability issue and second, whenever they closed the program and therefore stopped a session. In this way, we implemented an event-based design with two distinct events that were easy for participants to recognize. As we expected the diaries to be very low-level and perhaps even emotional, we thought that we would need a more reflective view on the issue from our participants and thus included the semi-structured interviews in the study design. Another important issue was the combination of diary logs and interaction logs. The interaction logs allowed us to see exactly which keywords users typed in and which functionalities of the tool they used; however we could not understand the purpose behind these actions. The diaries allowed us to capture the Why in much greater detail without having to ask the participant to report on every functional detail of the interaction with the system. In addition, the combination of all three methods also allowed for cross-validation and an assessment of the compliance. This becomes obvious in the example of the diary logs and the interaction logs. Some participants handed in quite a few diary sheets with many session reports; however, looking at the log file, we could see that they had used the system only the day before for a couple of hours. Thereby, the interaction logs allowed us to validate the diary logs and to see how compliant and truthful participants were. All participants were informed about the interaction logs before the study.

Table 6: Diary template (translated from German)

User-protocol
for **MedioVis**

ID:	Date:	Start-time:	End-time:

1. Please describe briefly for which purpose you have used MedioVis. You can mark tasks within the list or describe it in your own words. You may make multiple marks.

☐ Goal-directed search for a movie
☐ Explorative search/browsing
☐ Search for movie descriptions
☐ Search for movie ratings
☐ Search for movie trailers
☐ Search for people involved in a movie

2. Did you use links to external sources during your work with MedioVis (e.g., imdb.com)?
☐ No
☐ Yes, I used: _____

3. Did you have to rely on external sources that were not offered directly in MedioVis?
☐ No
☐ Yes, I used: _____

4. Did you encounter any problems during your work with MedioVis?
☐ No
☐ Yes, these were: _____

5. How much fun did you have using MedioVis?

☐ Very little ☐ little ☐ average ☐ A good deal ☐ a lot

1. Please describe all the things that you liked and disliked:

_____
_____

3.3.3 Results and Discussion

One rather problematic issue with diary studies (although seldom reported in the HCI literature, for some reason) is the increased chance for participant drop-outs. In many cases, participants are not really aware of the burden of participating in a study for several weeks or of keeping a diary every day or several times a day. Therefore, they underestimate the effort needed and after a few days they realize that they are unable to further participate in the study. In our case, three of the eight participants had dropped out by the beginning of the second week. There are several possibilities for decreasing the chance of such drop-outs. First, it is possible to split the financial reward into parts so that participants are encouraged to continue in order to get the full amount. Second, the relationship between researcher and participant plays a large role. If the researcher is helpful and supportive all the time and also shows that he or she cares about everything the participant records, this can help a great deal in increasing the intrinsic motivation “to be a good participant.” However, there is also a potential downside to this: Some participants may misinterpret “being a good participant” with “saying what the researcher wants to hear.” Therefore, it is essential that the researcher does not take sides and emphasizes that he or she does not have a preferred view on the study subject (e.g., which of the tools is “supposed” to perform better).

Our high-level results showed that both visualization techniques worked reasonably well; two participants even asked if they could continue to use them despite some technical bugs that limited the usability. Switching to the second system after one week was perceived as quite an easy undertaking, mainly due to the similar interaction concept applied in both systems when accessing an information object. We found that the HyperGrid was better suited to searching for one specific object, while the HyperScatter provided ways to compare information objects and to look for interesting clusters and correlations. In both cases, users took advantage of the browsing possibilities and especially liked the integration of external web services such as youtube.com and imdb.com. Because of this function, they could focus on a single system and did not have to constantly switch between webpages with different interfaces. Interestingly, two

of our participants came to the conclusion that it would be beneficial to combine the two approaches (HyperGrid & HyperScatter) into one system; the other three were also positive about this possibility when asked.

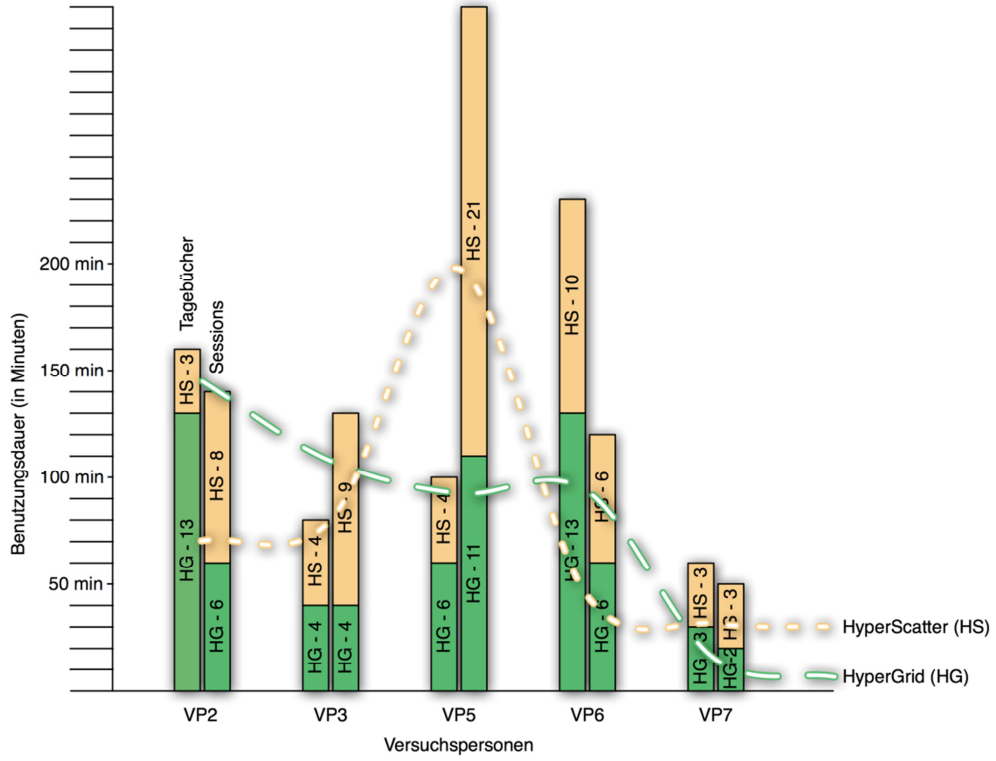


Figure 25: Relationship between diary entries (left bars/participant), session numbers (right bars/participant) and usage duration (y-axis + dotted lines)

Examining the detailed results from a methodological perspective, we can see several things. First, Figure 25 neatly displays the relationship between diary entries and interaction logs. For example, participant 6 (VP6) completed a total of 23 diaries but only 12 logged sessions were recorded. Not all of them could be explained by multiple diary reports per session. While people marked “explorative search” as their purpose only in nine diary reports, the possibility for browsing was reported as a major benefit in the interviews. The diary reports also revealed that people often used the integration of external web sources (in 17 of 39 logs for the HyperGrid and 10 of 24 for the HyperScatter), which could be confirmed by analyzing the interaction logs. During interviews, participants stated that they really appreciated the possibility to access external data sources such as Google or imdb.com directly from the application. People also

stated that they liked the possibility to freely assign metadata to a column or the axes. The log files confirmed that people not only liked this function but also used it heavily (in 25 out of 47 sessions total).

Overall, the triangulation paid off extremely well. It became quite obvious that the three data sources each revealed a different kind of information. The interviews revealed overall impressions of the participants and allowed them to elaborate on certain problems and express suggestions for improvements in detail. The diaries, on the other hand, allowed us to get a better understanding of how and why the system was used, with detailed information about purposes of usage and specific problems and how these affected the overall rating of the system at that time. Furthermore, this information provided helpful during the interviews for elaboration. The interaction logs were mainly used to check how the qualitative expressions were reflected in real usage of the system and thereby proved to be very valuable to judge compliance, i.e., whether there was some “substance” to the talking (to see whether people really used a function they praised in an interview, for example). However, there was still room for improvement. When evaluating software running on a PC, it seems preferable to directly integrate the diary functionality in digital form. In this way, participants could be reminded to fill out a diary whenever they closed the system. In addition, an integrated on-demand diary function could be used for usability issues. This could also trigger a screenshot, which would help during the analysis process. Unfortunately, integration of such an electronic diary directly into an application requires development resources that might not be available. A pen and paper diary proved to be a very cost-effective and informative solution.

The PocketBee approach presented in the next section could also be a very effective solution.

3.4 PocketBee – A Multimodal Diary and ESM Tool for Longitudinal Field Research

Parts of this chapter were published in (Gerken, Dierdorf, Schmid, Sautner, & Reiterer, 2010b)

As technology becomes increasingly ubiquitous in the modern world, HCI research is thus ever more interested in investigating how people deal with technology in the wild. Consequently, the need for diary and ESM methods has increased, since they are capable of capturing this interaction. Technology itself has been a helping hand, supporting both the researcher and the participant: the introduction of electronic diaries, previously based on PDAs and now on smart phones, allows integration of the method more seamlessly into the daily lives of participants. In this chapter we present PocketBee (see Figure 26), a multi-modal diary tool that allows participants to gather data in multiple ways on Android-based smart phones; it also allows researchers to access this data immediately via a web-based control center and react to it accordingly, e.g., by sending out specific tasks or questionnaires. PocketBee integrates an easy-to-use client user interface that reduces the burden on the participant while maintaining a high degree of flexibility with respect to the method and the possibility to capture in-depth data. We furthermore discuss several research questions that illustrate the importance of the tool for both the diary method and for ESM.

3.4.1 Introduction and Research Questions/Design Goals

The design of the PocketBee diary tool was motivated and guided from two perspectives: HCI research in general and the automotive sector (for which we designed and developed the PocketBee diary) in particular. We closely collaborated with a well-known manufacturer of luxury automobiles with a high percentage of discerning customers. In such a case, a profound familiarity with modern technology such as smart phones cannot be taken as a given. Comfort is one of the most important aspects when these customers are deciding to purchase a vehicle. This stresses the importance of an easy-to-use yet powerful and flexible user interface even more.

Figure 26: PocketBee running on a Motorola Milestone

Early electronic diary or ESM tools focused on simply providing questionnaires on a PDA (Barrett & Barrett, 2001), while current approaches have focused especially on the integration of sensor data to better support ESM (e.g., (Froehlich, Chen, Consolvo, Harrison, & Landay, 2007)), multiple modalities to enrich the data-gathering process (e.g., (Carter & Mankoff, 2005), (Jain, 2010)), or the integrated testing of mobile device applications (Carter, Mankoff, & Heer, 2007). The existing tools seem to have focused on functionality and extensibility but not so much on the design of the client user interface itself, as it has been merely discussed in the respective papers. However, an electronic device does not magically reduce the participants' burden of collecting data; it might even increase the burden for some users who are not familiar with smart-phone technology. In the following sections, we will first present our research questions for the user interface design and go on to discuss the user interface concepts by illustrating an upcoming study in the automotive sector.

3.4.1.1 Research Questions

Khan et al. (Khan, Markopoulos, & Eggen, 2009) conducted an analysis of current experience sampling tools, deriving some requirements for future tools, such as multi-modality and instant synchronization. While we agree with most of these and they influenced our choice of design goals and research questions, the underlying motivation for them is often much more technical or practical in terms of what might be useful for the researcher. As previously stated, however, we will focus on the research questions from a user interface design perspective and the methodological benefits that can thereby be achieved.

- *Reduce the burden on the participant:* We think this should be one of the primary goals of any diary or ESM tool. In principle, this starts the portability of the device – in a best-case scenario, this can be achieved by using either the user's own smart phone or by temporarily replacing it. In addition, it should provide the most convenient means (as in (Jain, 2010)) of data-gathering for any situation as well as being easy to use. This means not only that all functionality can be easily accessed but also that the UI is designed in such a way that the user is reminded of his or her tasks without having to look them up.

- *Link the data-gathering closer in time to events:* In order to reduce retrospective memory effects, data-gathering should occur soon after the events in question. In the case of ESM, this is essential (e.g., as in (Froehlich, Chen, Consolvo, Harrison, & Landay, 2007)), but also for diaries we find that time- and sensor-based designs should be supported. This basically means that 1) the diary must be carried along at all times and 2) either the device must prompt the participant or that the participant must be instructed to capture data instantaneously when certain events take place.

- *Increase the quality and depth of collected data:* Early electronic diaries only incorporated simple questionnaires, which made it difficult to benefit from the "*in situ*" quality of the diary. We think it is essential that a diary allows the capture of rich data that is both appropriate to the situation and the partici-

pant (as stated above). This should include both manually-triggered data-gathering as well as automatically triggered or logged data. This also means that both qualitative and quantitative data-gathering should be integrated.

- *Increase the strength of the bond between researcher and participant:* Motivation is a serious issue in diary studies. We think that having a direct communication channel provides participants with the assurance that their comments and feedback are being recognized, making it valuable and worthwhile to continue contributing. This also means that researchers should be able to examine and analyze diary material before meeting with participants for an interview, supporting in essence the design of an elicitation diary.

- *Bridging the gap between ESM and diaries:* The tool should be designed in such a way that it can be adapted to either focus on experience sampling by triggering questions or tasks at certain events (as in ESM) or by manual user input (as in traditional diaries). Overall, it should support any combination of condition-based and human recognition-based designs. It should be possible to adjust to changes during the study and react to user data (e.g., by adding additional tasks or questionnaires remotely).

3.4.2 Related Work

In this section, we would like to provide a brief overview of existing (electronic) diary/ESM tools and their characteristics. We will focus especially on the kinds of research designs the tools support and the ways researchers and participants can interact with them. The traditional physical form for recording information in a diary is pen and paper. Ideally, the paper would already be bound in a booklet in A4 or A5 size, so that participants would not have to look after single sheets of paper. The A5 size adds mobility; if people are asked to carry the diary with them, the A5 size has some advantages. Such a booklet should also contain an instructions page. Each page is assigned to one diary entry to facilitate analysis. In the case of time-based designs with random intervals, as in ESM studies, an additional signaling device is needed. This augments the pa-

per diary with an electronic device such as a pager or alarm clock which notifies participants when a diary entry is due. While such a paper diary can be nicely integrated into the personal environment of participants, it does have certain drawbacks. First of all, it is often difficult to carry around, even at A5 size. Especially for time-based designs at random intervals, this creates a problem, as the diary may not be at hand when needed. In addition, researchers have no guarantee of when diary entries are actually completed; again, this is especially problematic in random-interval designs but also in general. If participants cannot be trusted to complete the entries at the given times or right after events, a recall bias may be introduced, making interpretation of the diary data difficult. However, it is true that paper diaries are relatively easy to set up and maintain and can be provided to almost everybody without a “technical” instruction.

ESP (Barrett & Barrett, 2001) was probably one of the first ESM tools on a mobile device (a Palm). In comparison to more recent tools, it had very basic functionality, focusing on time-based random schedules with questionnaire items as response types on the device. The configuration was possible through configuration files. As it is Open Source, ESP has had a long history of usage in studies in a variety of fields, including the social sciences, psychology, and HCI. Momento (Carter, Mankoff, & Heer, 2007) was one of the first tools that emerged specifically from the HCI area. It used the more modern Windows Mobile OS, which allowed it to be deployed on a variety of devices. The functionality focused very much on questionnaire items and a reporting functionality that allowed participants to send text and images to the researcher via SMS/MMS. The tool also provided a monitoring application that allowed researchers to visualize and analyze data on the fly. Although it is also Open Source, the development of the tool seems to have stopped a few years back. MyExperience (Froehlich, Chen, Consolvo, Harrison, & Landay, 2007) and Xensor (Hofte, 2007) seek to make use of all of the sensing capabilities of modern smart phones in order to enhance ESM studies. They allow the combination of automatic logging of sensors to gain objective data as well as subjective ESM-like user responses. Both Xensor and MyExperience provide a very modular architecture that allows easy integration of new events or sensors for developers. MyExperience allows researchers to define events and conditions via XML. For

both systems, recorded data is automatically transferred to a server when an internet connection is available. While the presented tools mainly have focused on supporting ESM studies, others have taken the logging approach even further. Recon (Jensen, 2009) allows the application to attach to a host application (by source-code integration alone) with the goal of obtaining objective usage logs for evaluating the host application itself. EmotionSense (Rachuri, et al., 2010), a life-logging tool, is similar to SenseCam (Hodges, et al., 2006), except with the goal of researchers being able to automatically recognize emotions.

A few tools have also tried specifically to enhance the diary method, thereby focusing on the ability of the participant to record subjective data without being prompted (e.g., (Hammer, Leichtenstern, & André, 2010), (Jain, 2010), (Khan V. J., Markopoulos, Eggen, Ijsselsteijn, & Ruyter, de, 2008)). ReconExp (Khan V. J., Markopoulos, Eggen, Ijsselsteijn, & Ruyter, de, 2008) combines the diary approach with the day-reconstruction method to increase the validity of diary data by conducting tool-supported remote interviews with participants at the end of each day. InfoPal (Jain, 2010) provides a rich set of data-gathering possibilities and a graphical user interface on the mobile device that allows participants to create a multi-modal diary entry (e.g., text, voice, video), including combinations of modalities.

PocketBee differs from existing tools, as our goal from the beginning has been the combination of diary and ESM designs. This is reflected in the event architecture and the design of the user interface on the mobile device and for the researcher. While other tools may be able to e.g. adapt a diary behavior through the “misuse” of questionnaire items, PocketBee has been designed to support this whole range of research designs natively, as we will illustrate in the following sections. Furthermore, we focused on the ease of use and learnability of the client user interface. We introduced the notion of Core Questions to help participants remember their data-gathering task. PocketBee is being developed for the Android OS and uses a client-server architecture. The mobile devices communicate with the server via their data connectivity. The researcher can define and manage a study remotely; settings are automatically transmitted as an XML document to the mobile devices. These interpret the logic and are notified by a

push service of any changes the researcher may carry out. All material that has been uploaded from devices is accessible remotely in our web-based control center.

3.4.3 Event Architecture and Relationship to the Research Design Classification Scheme

For the PocketBee diary/ESM tool, we designed an event-architecture (Figure 27) that supports all different study designs and is flexible enough to also handle all kinds of combinations. In principle, we rely on a three-level architecture with *sensors*, *conditions*, and *actions*. Conditions evaluate sensor values and trigger actions. Multiple conditions can be combined into a *condition chain*. The design for our architecture was inspired by the excellent work of the MyExperience project (Froehlich, Chen, Consolvo, Harrison, & Landay, 2007). Our contribution thus is based on advancements with regard to the flexibility and the complete range of study designs we are able to support, including human recognition-based and mixed designs. We will illustrate this by establishing a direct association to the proposed classification of study designs (see Section 3.1.4).

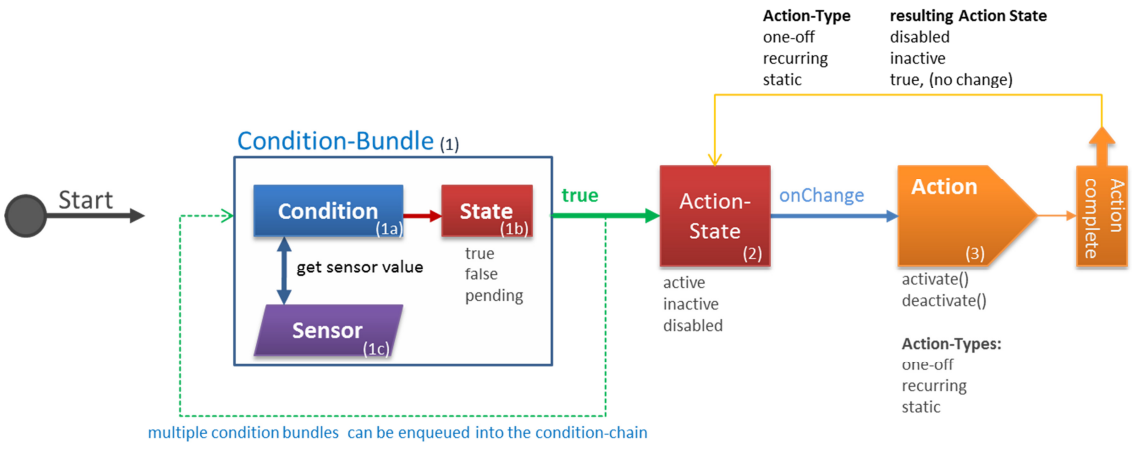


Figure 27: Event-architecture of PocketBee for diary/ESM study designs

3.4.3.1 Actions

Actions (Figure 27– (3)) represent the top level of the architecture and initiate the data-gathering/response prompts that ask participants to input data. As we want to support both manual and automatic data-gathering, we distinguish between (automatic) *system actions* and (manual) *user actions.*

A user action can be distinguished into three types: *one-off, recurring, and static user actions*. For human recognition-based designs, we require the possibility of manually recording data whenever participants feel it is necessary. In a pen and paper diary, one would use a new sheet of paper for every diary entry; in an electronic diary, we need a new instantiation of a data recording screen every time the participant decides to record data. However, the possibility for the user to start recording data must be present at all times. Therefore, we call such actions *static*. For condition-based designs, we need two different kinds of manual data-gathering. First, we need the possibility of recording singular event-conditions (i.e., events that only happen once, such as a questionnaire to assess overall impressions at the end of the study). We call these types of actions *one-off actions.* Second, we also need a possibility for recurring data-entry, such as daily questionnaires. In this case, we need a re-instantiation of the data-recording screen each time, similar to the static user actions. The difference here is that the condition chain must be reevaluated for such an action to recur. Therefore, we call these user-actions *recurring. System actions* can be used to start internal system processes that do not require any user input. For example, an action could start logging a particular sensor (e.g., a camera or a heart rate sensor) for later analysis.

Each action can be assigned to one or multiple participants/devices that are currently registered in the system. This allows the researcher to define individual study designs as well as to use a “one size fits all” approach. Furthermore, creating groups allows easy manipulation of within- or between-subjects studies.

3.4.3.2 Action-State

Every action is controlled by its *action-state* (Figure 27 – (2)). If this state is set to active, the action is started. In case of user actions, the action is now waiting for user input either by using notification mechanisms (e.g., ringing or vibration of the device) or simply via a visible UI element on the device that allows the participant to record data. Upon completion of an action, it notifies its action-state, depending on its type. A *recurring action* sets the action-state to inactive. It can be reactivated again if the condition chain becomes true once again. A *one-off action* disables the action-state and any preceding condition chain. This action will not be restarted again. A completed *static action*, as used for human recognition-based designs, however, does not change the state (thus, it stays active) but reinitiates itself so that new input can be received (such as a new diary entry).

3.4.3.3 Conditions

A condition (Figure 27 – (1a)) is a programmatic module that is closely coupled to a specific sensor. It periodically polls the sensor to get its current value and then checks this value against pre-defined value(s). The comparison of polled sensor data and pre-defined values can return either true or false. Upon a change in the return value, the condition sets a corresponding *condition state* (Figure 27 – (1b)) accordingly to either true or false. Multiple conditions can be combined together and interlinked with a logical AND. They are also connected to an action via the action state. The position within the chain and the current state is stored in a tree structure. Upon a state change, the AND combination of all condition-states that are linked together is checked with the tree structure and accordingly the adequate action state is set to active (if all states are true) or inactive (if one or more states are false). We call the combination of sensor, condition, and condition state a *condition bundle* (Figure 27 – (1)).

3.4.3.4 Sensors

A sensor (Figure 27 – (1c)) on the architecture level is an abstraction class to either a physical sensor or a virtual sensor. For example “time” is treated as a virtual sensor, and the corresponding sensor class can poll the sensor to re-

ceive the current time. GPS, on the other hand, is a physical sensor in the device and the abstraction class can poll this sensor to receive the current location coordinates. Every sensor class stores the values from the latest polling in addition to a timestamp. The condition can thereby evaluate whether it can use the latest value (without needing to invoke the GPS, for example) or whether the sensor class must poll the sensor for the most current value. In this way, we can preserve battery life in the case of multiple conditions using the same sensor.

Table 7: This table shows the relationship between different study designs and the system architecture with examples for study designs. Combinations of designs are not included here, but are supported by the architecture.

Diary/ESM Design		Sensor	Conditions	User- Actions (manual)			System-Actions (automatic)
				one-off	recurring	static	
Condition-based	Time-based	Time	Time-schedule (random or fixed)	Example: Pop-Up Screen (modal dialog) asking the participant to report her current emotional state on an affect grid. Condition: at 6pm on the 2nd of February 2011,	Example: Notification about a new task, asking the participant to report her current emotional state on an affect grid. Condition: recurring randomly once every 2 hours	Not applicable	Example: Log the heart rate sensor. condition: between 6 and 8pm
	Sensor-based	Any physical or virtual sensor	Sensor-conditional event (fixed)	Example: Notification about a new questionnaire, asking the participant to report about a day. Condition: as soon as the participant arrives at home (GPS/location range sensor).	Example: Pop-up of a new questionnaire (modal dialog) asking the participant what caused a physical exertion. Condition: Every time the heart-rate sensor has recognized an average heart rate above 130 for more than 15 minutes;	Not applicable	Example: Poll all consecutive conditions to update their sensor values. Condition: As soon as heart rate is above 150 for 5 minutes.
Human-recognition based	Researcher defined events	-	-	Not applicable	Not applicable	Example: Asking a participant to record every barrier she can find for physically disabled people in the university building	Not applicable
	Participant-defined events	-	-	Not applicable	Not applicable	Example: Asking the participant to record any problem she has with the current CMS software.	Not applicable

3.4.3.5 Specifics

In Table 7 we provide an overview of the relationship between the classification of study designs and the event architecture. All four study designs are supported and, although not visible in the table, combinations of designs are also possible. For example, we can have a static action visible to the participant only

between 6:00pm and 8:00pm, thus combining a human recognition-based design with a condition-based (time-based) design. In addition to the basic structure of the architecture with actions, conditions, and sensors, we discuss some further specifics in the following paragraphs:

- Boolean Concatenation

  To enable the researcher to define more complex statements, we introduce a Boolean filter, which allows the researcher to treat a "false" condition-state as "true." Imagine a location condition: it would be valuable to define different actions depending on whether a person is within a certain location range or outside of it. Instead of having to define separate condition-action chains, the Boolean filter allows the researcher to use both outcomes of the condition (participant within or outside of the specified range) for different actions.

- Sampling frequency

  In an ideal research design, the system would know all current sensor values at all times. However, for some sensors, this would simply result in a high battery drain (e.g., in the case of GPS). Therefore, each condition has a researcher-definable sampling frequency for polling the attached sensor. An overly large sampling frequency (e.g., every six hours) could lead to events being missed completely, whereas an overly small sampling frequency will inevitably lead to a dead battery in short work. Therefore, this value must be chosen with consideration of both the research question and the battery drain. As already discussed, each sensor-class stores the latest values and timestamps. If a condition polls the sensor-class, the timestamp of the latest value is first compared to the sampling frequency. A new value is only actively retrieved if the timestamp is too old (i.e., not within the sampling frequency). In case of multiple conditions making use of the same sensor, we avoid unnecessarily multiple polling of this sensor.

- Pending state

  In addition to a condition-state being true or false, it can also be set to *pending.* To illustrate this idea, we will use a specific example: In the case of

time-based conditions, it can be essential that an action is performed immediately on time and not simply within the boundaries of the sampling frequency. The Android OS features an alarm manager that can notify the system as soon as a certain time is reached. The *pending-state* allows us to have the alarm manager directly set the state to *true* and initiate an active polling of all other connected condition bundles. In general, the pending-state allows external actors such as the researcher to intervene in certain pre-defined situations. By introducing the pending states, we can model these situations in the same way to on-device conditions.

- Actions as conditions

  An action can also function as a condition for further condition-action chains. To illustrate the need for this, consider branching in questionnaires. Based on specific (and predictable) user input, the system presents further tasks. For example, a certain questionnaire should only be presented if another has already been completed, or the system should present a questionnaire only after a participant has taken a photograph. By using our architecture, we provide flexible handling of such situations with the ability to patch different data-gathering or response types.

3.4.4 User Interface Design

In the following section, we will illustrate how PocketBee works and how the user interface is designed to address the research questions.[7] To this end, we will present a scenario of use.

[7] A short video demonstrating the PocketBee tool can be accessed online: http://www.vimeo.com/13397614.

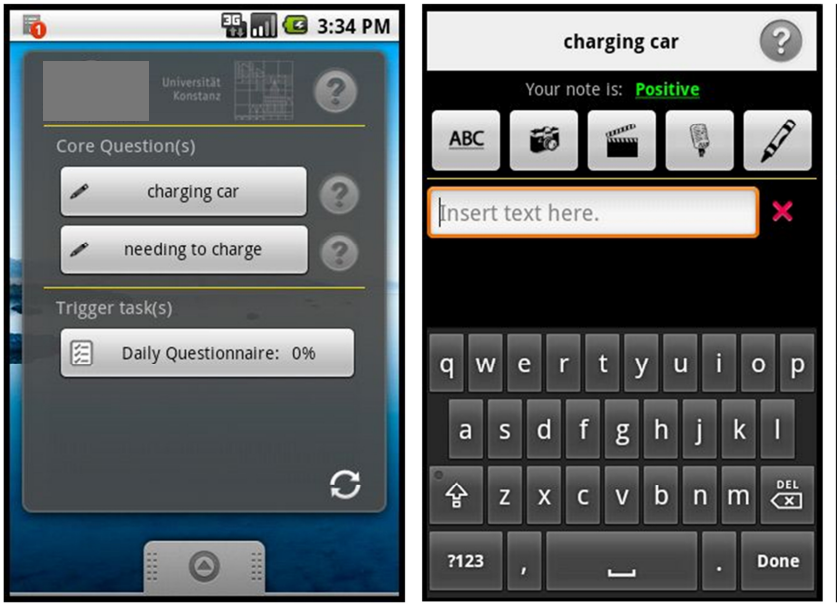


Figure 28: left: Home-Screen Widget with 2 core-questions and a questionnaire (lower part), right: diary entry form (empty)

The PocketBee client diary application is written in Java/Android. It directly integrates drawing and text notes modalities and additionally integrates the Android's internal camera, video, and voice recording applications seamlessly into the user interface. The user interface consists of a home-screen widget and the data-gathering application (see Figure 28). The widget allows the participant to use the mobile device while having constant access to the diary application. It provides entry points for the user and acts as a constant reminder of any pending tasks. Essentially, the widget supports both human recognition-based diary designs and condition-based designs. The upper part is reserved for *core questions*, such as are applied in a human recognition-based design. These core questions serve as visual and cognitive triggers; the user can simply wait for these events to happen and is constantly reminded to "get triggered" by them. Upon selection of a core question, the interface allows the participant to compose a diary entry out of several notes and multiple modalities, including text, photograph, video, voice, and drawing. Upon saving, a diary entry is immediately sent to the server in the background. By providing different modalities for data-gathering, PocketBee increases flexibility for the participants, as they can simply select the most convenient one. The bottom part of the UI is reserved for condition-based designs, which are reflected in PocketBee by task and ques-

tionnaire actions. Alternatively, the researcher can define these actions to be displayed as a modal *pop-up* dialog that is displayed immediately after conditions are met. For the questionnaire, we support different item types, including rating scales, ranking, open-ended questions, sliders, and the integration of additional modalities (e.g., voice). Participants are always notified of new actions by the internal Android notification system, which then shows the action icon in the top status bar.

The interface concept supports all of the types of research designs that we have presented in Chapter 3.1.4. Furthermore, the widget concept allows the participant to be aware of actions even in a non-modal dialog setting, which might be most appropriate in human recognition-based designs.

3.4.4.1 Scenario of Use

Electrically powered cars are not only environmentally conscious but add to the customers' mobility and flexibility. Instead of having to rely on fixed gas stations, any power outlet can become a source for recharging. While little is known about how practical this might be, investigating it is difficult, to say at least. Direct observation is scarcely possible, since the car is a very private environment. Using interviews, retrospective effects might hide the little hurdles one must master during the charging process. We will outline in the following section how PocketBee can support such a study. The PocketBee client's user interface consists of a home-screen widget (see Figure 28, left) and the diary application itself (see Figure 28, right). The widget allows the participant to use the phone itself while having constant access to the diary application. It provides the entry points for the user and is a constant reminder of any pending tasks. Essentially, the widget supports both condition-based and human-recognition based designs. The upper part is reserved for what we call *core questions*. As it can be a mental burden for participants to constantly think about whether they should record a diary entry during any given situation, these core questions serve as visual and cognitive triggers to reduce this burden. In our scenario of use, these are a "charging car" event and a "needing to charge" event. The user can simply wait for these events to happen and is constantly reminded to "get triggered" by them. Such a human recognition-based diary design also allows researchers to

couple the diary entries closer to the events that need to be reported, as they motivate an instant capturing. By tapping on a core question, a diary entry is created that can then be enriched with data (see Figure 29, left). Let us assume that our participant Sarah is about to charge her car. The interface allows her to compose a diary entry out of several notes. To begin with, she might want to simply write a text note, that she is about to charge the car at a friend's place. During charging, the display in the car tells her how long it takes to fully charge the car. She takes a picture of the display and adds a textual note. She would like the researchers to know that she would like to enter the distance she wants to drive so that she knows how long to charge for a specific ride. She then saves the diary entry; it is now immediately sent to the server in the background, together with her current geo-location (if she has agreed to this prior to the study).

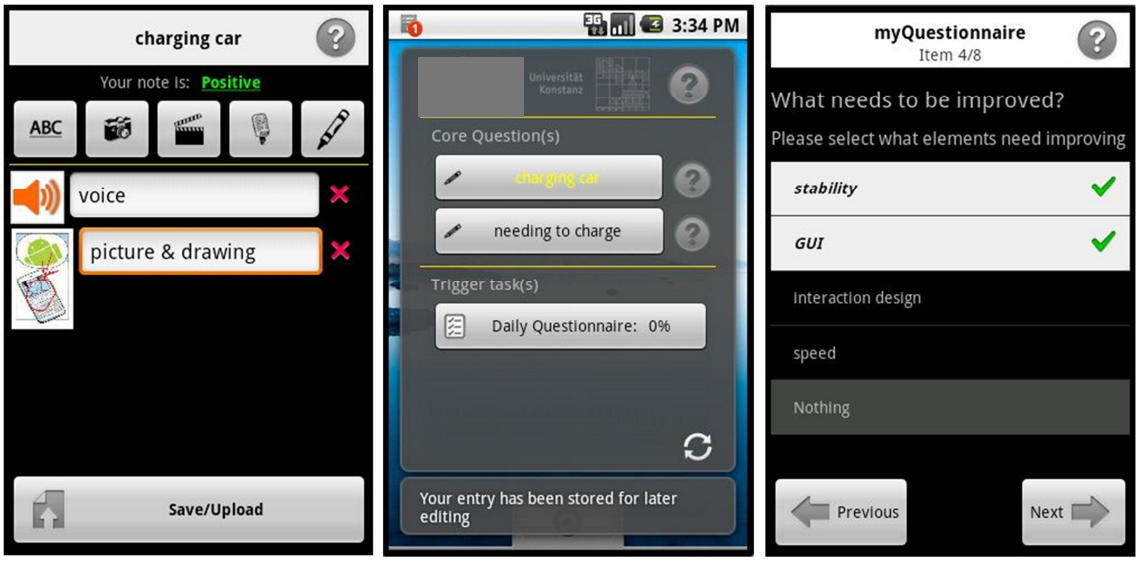


Figure 29: left: diary form with two entries (voice and drawing), middle: temporary postponed entry, right: questionnaire item

Later on, she gets another idea: the car should send her a text message as soon as the charging is complete. She quickly records an audio note while walking to the car to check the status by herself. By providing these different modalities for data-gathering, PocketBee reduces the burden on the participants, as they can just select the most convenient one. By allowing the composition of several modalities in one entry, we seek to provide rich and in-depth data.

Moreover, the GPS location can help during an additional retrospective interview to remind the participant of this particular situation for further discussion. The researcher, on the other hand, has immediate access to the diary entry via the control center (see Figure 30) or as soon as the device has a network connection (WiFi or GSM/3G). This allows the researcher to 1) start the data analysis right away, 2) prepare the data for an interview session, and 3) react to the data. We currently provide a basic list-like view for the entries that can be sorted and filtered by several criteria (e.g., core question, participant, etc.) as well as exported for data analysis (with MS Excel, for example). In order to react to the data, the researcher can modify existing or create new core questions as well as create additional tasks and questionnaires individually for each participant. The last two reside on the lower part of the home-screen widget. Tasks are meant to provide specific instructions, such as “please take a picture of the power cable,” allowing the researcher to interact more closely with the participant, tightening the bond between the two as the latter receives direct feedback on his or her actions. This will also help to increase the motivation for continuous use of the diary. Questionnaires can be designed in an XML template that provides several different question types for most necessities, such as multiple selections, rating scales, or open-ended questions (see Figure 29, right). This last option also allows the participant to record voice instead of typing text. The template allows branching questions as well as forced or optional questions.

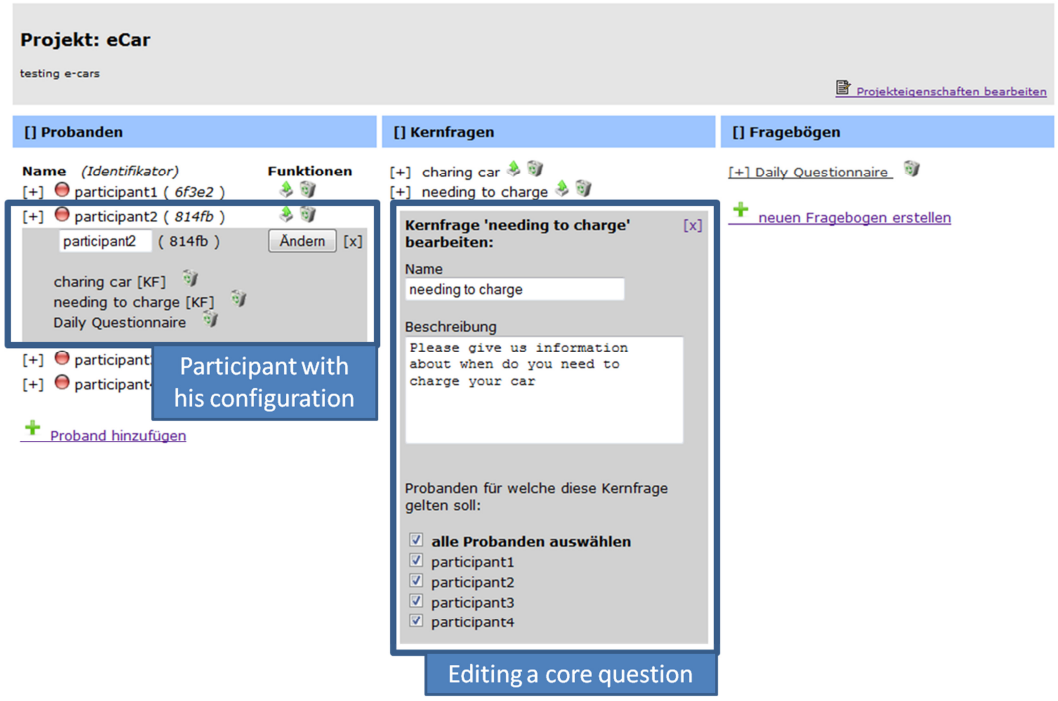


Figure 30: Web-based Control Center

Our participant Sarah comes home again. As on every evening since the study started, the device notifies her with two short beeps that the daily questionnaire is now available, asking her about the mileage she drove today, how she rates the ease of use of the charging device, and for additional feedback.

3.4.5 Implementation[8]

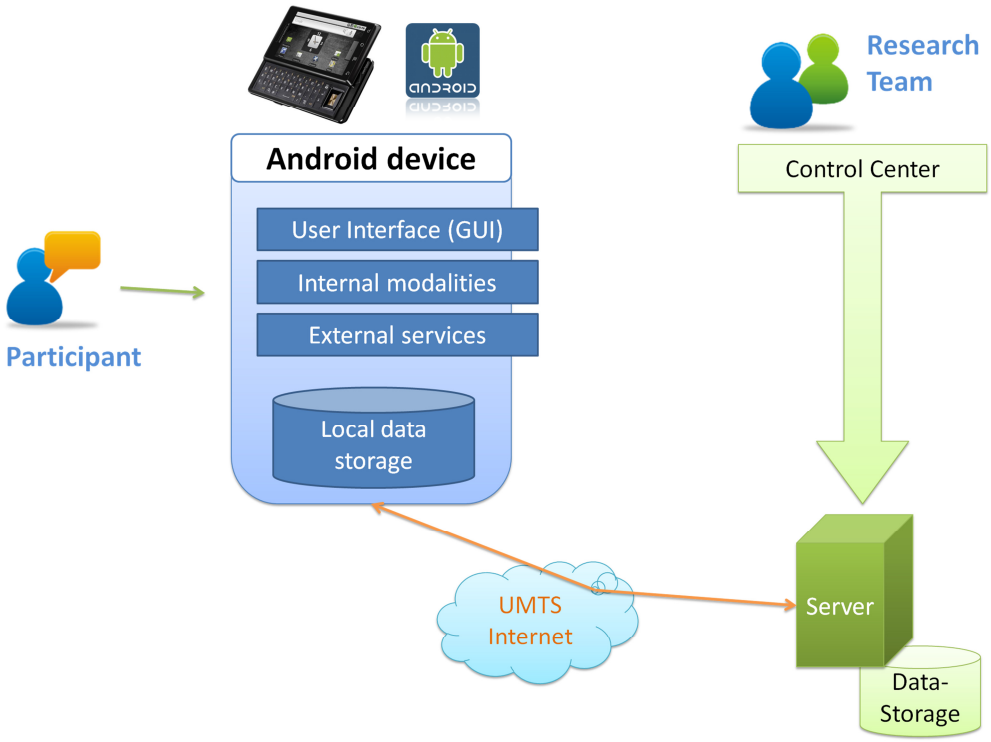


Figure 31: Schematic view of the PocketBee system

PocketBee is a distributed system including a mobile device (the bee), a server (the hive), and a control center for remote access to the server (for the bee-keeper, see Figure 31). PocketBee runs on every current Android-based phone (Android 2.01 and up). Given the steadily rising market share of Android devices,[9] this increases the chances of allowing the user to use their own device for the study. The researcher can both set up a study and manage multiple projects within the control center without having to touch the mobile device during the runtime of the study. The client diary application is written in Java/Android. It directly integrates drawing and text notes modalities and additionally integrates

[8] More details about the implementation can be found in the Master theses by Stefan Dierdorf (Dierdorf, 2011) and Patric Schmid (Schmidt, 2011). The implementation itself was not part of the author's work for this thesis.

[9] See: http://www.gartner.com/it/page.jsp?id=1372013 increase within the last year from 1.6% to 9.6% (checked on 30.07.2011)

the Android's internal camera, video, and voice recording application seamlessly into the user interface. Each diary entry is automatically tagged with the corresponding core question and the GPS location. If no network is available, a queue holds the diary entries until submission to the server is possible. In addition, the client attempts to contact the server every x minutes (default: 30). If successful, the client automatically asks for any updates available (e.g., new or modified tasks) and the server responds accordingly. The server hosts a MySQL database and a web server. All communications are handled via PHP scripts. Currently we support time-based triggers for tasks and questionnaires, which can be specified in the control center, similar to setting up events in a calendar application (e.g., daily questionnaire from 6:00pm to 10:00pm). The infrastructure, however, is built in a way that supports the full event architecture presented in Chapter 3.4.3. As a result, PocketBee allows the researcher nearly endless possibilities to create combinations of diary and ESM-like events and triggers for both automatic as well as manual data-gathering.

3.4.6 User Studies

In the following section, we will describe two studies in which PocketBee was used. This section will focus on the general research design applied and the usability issues we encountered with the PocketBee client UI. We think the latter are interesting as they provide some general insights for the future design of diary and ESM tools.

3.4.6.1 Well-being in a Car

We developed PocketBee in collaboration with the customer research division of a large automotive company. Their primary research goal is to ensure that customer input is systematically considered in important questions of research and advanced engineering. One research aspect here is comfort and well-being in a car. This is quite a difficult research area, as comfort and wellbeing in a car is not necessarily something people who own a car think about explicitly all the time. Rather, it is a more subtle factor adding up to the total experience of the product – or subtracting from it, if comfort needs are not met. Furthermore, it

cannot be directly observed, and asking people about it is also difficult, as it is not necessarily active in their conscious minds. A diary/ESM study allows the participants to think about this aspect in their natural environment. For our study, we selected 10 participants (between the ages of 28 and 52, 5 women), all customers and owners of cars the company produces. They received PocketBee on a Motorola Milestone (Droid) device with Android 2.0.1 OS for one week. We were interested in finding out what positive or negative experiences regarding well-being in a car they might encounter during the week and in which situations in general well-being is high. This last point was intended to provide further inspirations for car-designers.

Study Design

First and foremost, we applied a human recognition-based design, as we wanted participants to report and describe any incidence of well-being – both in direct relation to their car and in general. Participants were introduced to the core questions in an introductory interview session. Additionally, as we were interested in measuring the impact of visually displayed core questions on the PocketBee UI, only half of the participants were shown the questions explicitly on the UI. The other half simply had a “create diary entry”-action as a button on the home-screen widget.

We combined this human recognition-based design with a condition-based design, using fixed time intervals. To this end, we designed a questionnaire action that showed up every other evening at 7:00pm and disappeared again at 11:59pm if participants had not completed it before that time. This questionnaire asked participants to report about their day in relation to feelings of well-being in the car: for example, how much they were driving and which aspects of the car made them feel comfortable. Participants were notified of a new questionnaire by vibration and sound, and the new questionnaire also appeared on the home-screen widget. However, participants were free to choose when to complete it, as we were not interested in immediate reactions in this case. In addition, participants received a task action, which also appeared on the widget and notified participants with vibration and sound. The task asked participants to relax for a minute, sit down on a couch and think about what could improve their well-being

at that moment. The task appeared at 8:00pm in the evening (assuming that most participants would be able to sit down for a few minutes at that time), but participants were free to choose when to complete the task. As we were interested in a well-being moment, forcing immediate task completion could have had counterproductive effects. An additional task action at the end of the week explicitly asked people to record whether and how they had tried to increase their comfort within their car.

After one week, participants returned the devices and took part in an elicitation interview (Lazar, Feng, & Hochheiser, 2009). The PocketBee researcher interface allows the display of all data recorded by a participant, and additional comments may be entered. During the interview, the researcher asked the participant to elaborate on a pre-selected subset of submitted material to further validate the findings.

Overall, the study's design shows that one specific design was not used, but rather a combination. We think this is a sensible approach for most studies, and PocketBee provides the necessary flexibility to design and conduct such studies.

Results

We will briefly describe the results from a usability perspective. Overall, all participants reported that they got along very well with PocketBee and that they did not need any additional manual. When asked about what they liked most, 4 explicitly mentioned the easy-to-use interface and 6 the variety of possibilities to express oneself, i.e., the different modalities and their combinations. Criticism was based more on the device itself, which was not as fancy as an Apple iPhone. On the other hand, the physical keyboard proved to be an advantage, especially as some participants had no experience with touch-screen phones. Some mentioned that drawing on the screen was too imprecise, but still liked being able to draw on photographs to point out important aspects. With regard to the different modalities, most preferred text (5 participants), followed by photographs (4) and voice recording (3) (multiple selections were allowed). However, voice recording was also heavily disliked by some participants, as they did

not want to expose themselves in public; we find this interesting, as most people have no problem taking a phone call in public. We could identify a pattern in which using text messages and photographs were acceptable in public situations, while video and voice recording seemed awkward to the participants. From a design perspective it is interesting that most participants (8) claimed that they recognized new events by looking at the screen rather than by the vibration/sound notification (2). Therefore, we assume that having a widget concept that displays new actions in a prominent manner is an important aspect and probably more important than the notification by itself, as participants may either not recognize the notification or simply ignore it in inappropriate situations and later forget about it. Surprisingly, more than half of our participants did not find the additional device (in addition to their private cell phone) a burden. With Android's popularity increasing, it might be possible to install PocketBee on participants' own devices in the future. Regarding the visibility of core questions, we could not identify any significant differences between our two groups. Interestingly, when we asked them afterwards about their opinion of the other condition, they mostly preferred the one they had experienced, independent of whether their condition was the one with or without core-questions displayed on the screen. We will continue to research the effect of core questions in upcoming studies to clarify this issue. In summary, participants also appreciated the study design and the possibility to provide feedback at any time. Some even mentioned that they would have preferred a longer study duration to be able to provide even more feedback. Table 8 shows the number of diary entries provided by the participants. Regarding the research question, the study provided many suggestions for improving well-being in a car, from very specific criticisms (e.g., having more than one cup holder) to broader insights. Researchers from our collaboration partner were happy to have received a great deal of very concrete and rich suggestions and feedback that made it easier to derive concrete design suggestions for improving well-being in a car.

3.4.6.2 Accessibility in the University

In the second study, we were interested in identifying barriers for wheelchair users in the University of Konstanz building. To increase public awareness, we

asked non-disabled persons to look for these barriers and report them with PocketBee. Again, we selected 10 participants (between the ages of 21 and 30, 4 women).

Study Design

We chose a similar design to that of the well-being study with the focus on a human recognition-based design, asking the participants to identify and report on both barriers and accommodations for wheelchair users. In addition, a condition-based design was used with a fixed time schedule for both a task and a questionnaire. The task asked participants to find a route between two locations within the university building and to report all barriers along the way; the questionnaire asked for overall ratings regarding the accessibility of different parts of the building.

Results

Results are as encouraging as in the well-being study, with all participants claiming that they had no trouble using PocketBee. The criticisms we received were mostly device-based (e.g., difficulty typing with the keyboard, or the weak battery – in this study, we used a Samsung Galaxy S). As can be seen in Table 8 participants took many more photographs, which confirms our assumption that multiple modalities not only allow users to choose the most convenient one, but also the one which best suits the data-gathering request. As participants were asked to record physical barriers, photographs were best suited for this task. Similar to the other study, participants told us that they had no trouble recognizing new events and mostly did so by simply looking at the widget or the notification bar at the top (7 participants) instead of the sound/vibration (1). Overall, the study allowed us to identify a huge number of barriers, which we handed over to the University administration and which we hope will help to improve the mobility of wheelchair users.

Table 8: Overview of modalities used in the case studies

	Wheelchair study	Well-being study
Note entries/Person	17.4 (2.5/day)	14 (2/day)
Audio	3.7%	10%
Video	5.5%	2.4%
Photograph	**45.4%**	**14.6%**
Drawing	3.2%	3%
Text	**42.2%**	**70%**
Media (Audio, Video, Photograph, Drawing) with additional text input	**57% of all media entries**	**75% of all media entries**

3.4.7 Conceptual Design for a Researcher Interface to Control the Event-Architecture

We agree with Froehlich et al. (Froehlich, Chen, Consolvo, Harrison, & Landay, 2007), who stated that there are two distinct audiences for diary/ESM tools: the *participants*, who will use an interface on the device, and the *researcher*, who will have to design the study. In both cases, the respective UIs need to resemble the underlying architecture in order to support the entire range of study designs. Our efforts presented so far have covered only the client user interface, which has been implemented and evaluated in multiple studies. In this section, we will present the conceptual design for the PocketBee researcher UI, the PocketBee Designer. It incorporates a visual language based on a pipe/filter and zoomable UI (ZUI) concept. The interface concept was inspired by Squidy (König, Rädle, & Reiterer, 2010), a toolkit for modeling multi-modal interaction

techniques. Based on their discussion of the benefits of a zoomable user interface, the idea for the PocketBee Designer is to have a flexible interface that allows to visually model the event architecture as presented in 3.4.3. All screens in this section are visual mockups of the interface.

3.4.7.1 The PocketBee Designer UI

The task of the researcher is to set up a study and maintain it over the course of the study, and also to do the analysis. This conceptual design focuses on the set-up and maintenance parts. The set-up process usually includes configuration of devices and preparation of any data-gathering or response events. The web interface presented earlier allows the researcher to easily set up and manage a study remotely. However, this did not yet include the complexity of condition-based designs in general, as we have discussed here, focusing instead on simple time-based questionnaires and tasks. Handling condition-based designs (either alone or in combination with human recognition-based designs) makes the design of a study much more complex. To date, the MyExperience tool provides the most convenient approach by allowing the researcher to define a sensor-trigger-action logic in XML. However, as noted by (Khan, Markopoulos, & Eggen, 2009), this still requires the researcher to be tech-savvy, especially if initial configurations turn out to be faulty. We counter this problem by presenting a visual language approach that allows abstraction from detail if necessary, without compromising on the *ceiling* (Myers, Hudson, & Pausch, 2000) of the tool. Basically, we rely on a pipe-and-filter metaphor in combination with a ZUI. The pipe-and-filter metaphor visually resembles the condition chain, allowing researchers to easily combine conditions and link them to the actions. The ZUI allows smooth access to a detail view for the configuration of each condition or action.

3.4.7.2 The Pipe/Filter Metaphor and Semantic Zooming on the Canvas

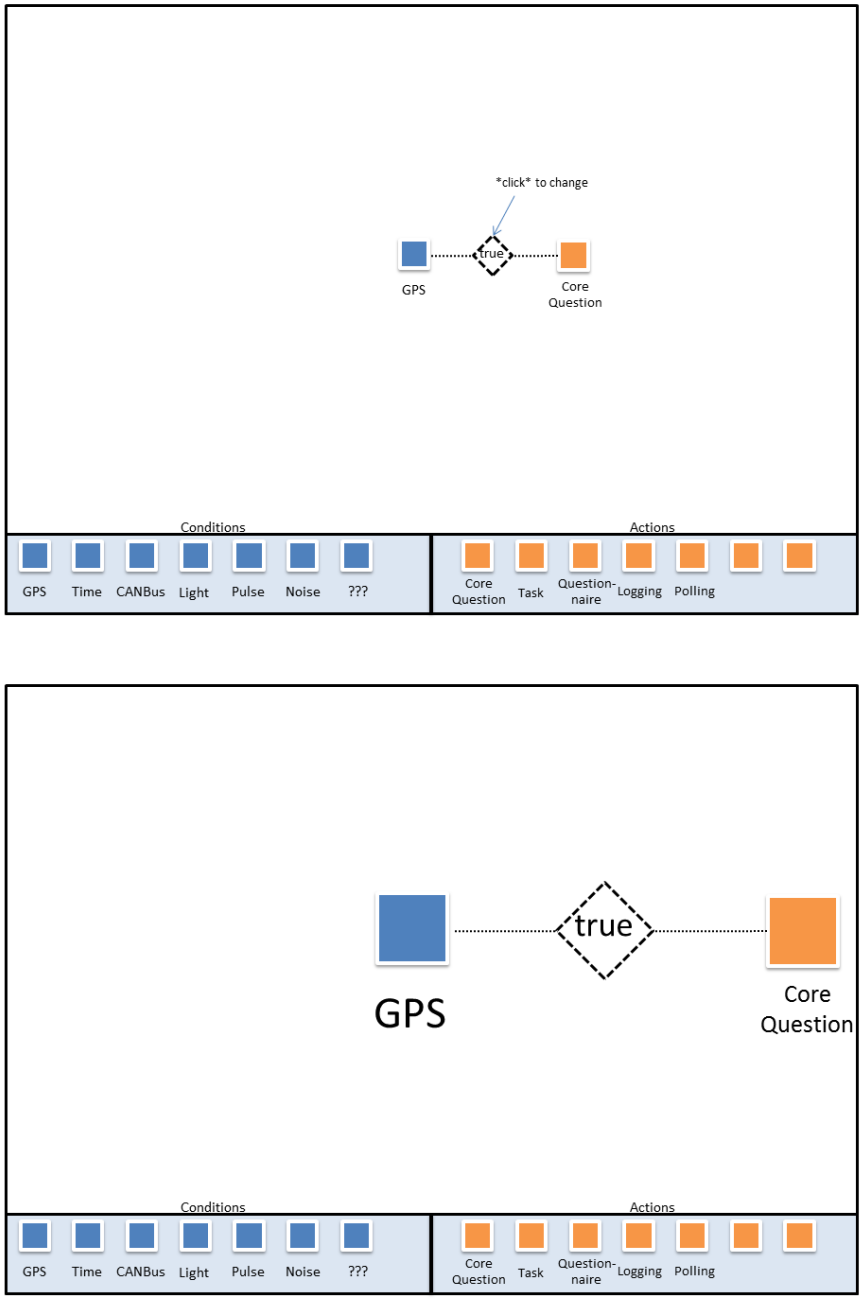


Figure 32: Pipe & Filter concept (left) and zoomable canvas (right)

A zoomable canvas serves as the interaction space for the researcher. Conditions and actions can be placed on the canvas via drag and drop and are represented as nodes. Visual links can be established between conditions and represent logical AND connections, resembling a data flow of true and false. Con-

necting one or multiple conditions with an action completes the condition chain (see Figure 32, top). By default, every action needs to be connected with at least one condition, which must trigger the activation of the action. For static actions in human recognition-based designs, this could be the time schedule of the study. For condition-based designs, more complex condition chains might be useful. Upon zooming into the canvas, more information and functionality is provided to the user. Each condition and action provides specific methods of configuration, which we will discuss in the respective sections below.

3.4.7.3 Toolbar

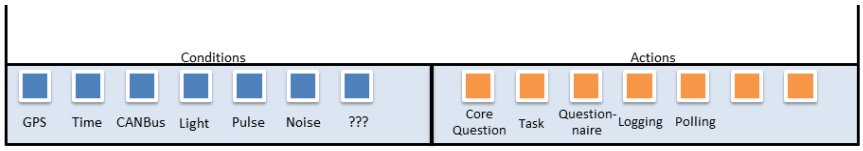


Figure 33: The toolbar with condition-objects on the left and action objects on the right

As every condition and action is implemented in a modular way, we aim to provide these modules as individual elements on the user interface, which can be easily extended if new modules are created. They are collected in a toolbar, which is accessible to the user in the bottom part of the UI. The user can simply drag and drop an action or condition on to the canvas.

3.4.7.4 The condition-object

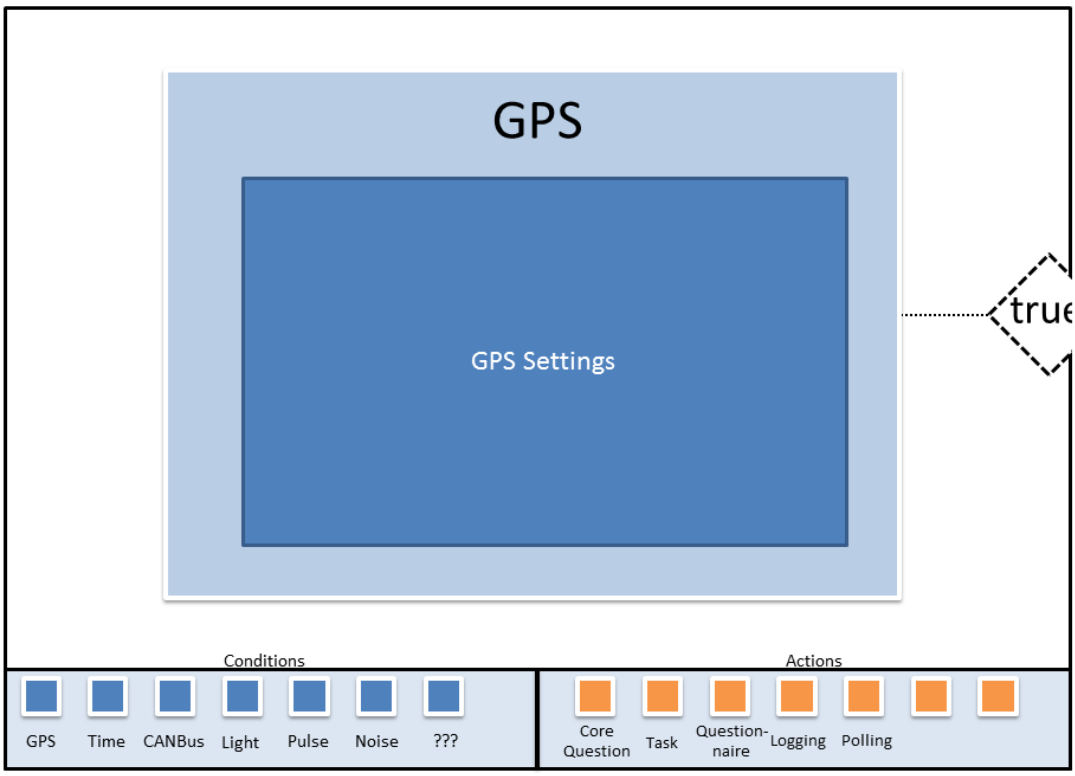


Figure 34: The GPS condition-object dialog appears after zooming into the node

From a user perspective, we think it is sensible to not separate sensors and conditions but instead to name conditions after the sensor they control. Semantic-zooming provides ways to configure the conditional handling of the sensor data. For example, a time-based sensor allows definition of the exact schedule. A location condition allows configuration of the exact position or range which should be used to trigger an action (see Figure 34, middle). Every condition module is integrated into the visual UI object. Upon zooming in, the user can choose between the various implemented modules available. For example, there might be two modules for time-based conditions, one for defining fixed schedules and one for random schedules. It is also possible to place the same condition-object multiple times on the canvas and to connect them in one condition chain.

3.4.7.5 The action-object

Action-objects behave as the condition-objects do. Each action-object needs at least one condition connected. On the user interface, zooming into an action provides additional functionality. Independent of the specific action at hand, the researcher can define the action type and for which participant or devices the action will be activated. Currently, we provide the following set of actions:

- *Tasks*

  A task asks the participant to fulfill a certain action. Bound to this action is a data-gathering or response activity. For example, a task could ask the participant to try out a certain functionality of a product and afterwards comment on it by recording a voice message. This action object thus allows the following factors to be specified: 1) A task instruction for the participants, 2) the modality for data-gathering they may use, 3) whether the dialog appears immediately as a modal dialog or within the PocketBee client UI so that the user can start the task whenever he or she is ready, and 4) an explanatory help dialog.

- *Core questions*

  For human recognition-based designs, we provide the notion of core questions. These are basically reminders of the pre-defined core situations in which participants should do the data gathering. Thus, instead of having a simple “create diary entry” button on the UI, the core question provides a visual reminder of what to look for. This action object allows the definition of 1) the core question itself as it appears on the UI, 2) a description of the core question, and 3) the modalities available to the participant for data-gathering. A core question always appears on the PocketBee Client UI and not as a modal dialog.

- *Questionnaires*

  Questionnaires can be used to serialize several questions, requiring the researcher to design the questionnaire. Our user interface allows a smooth transition between designing the event conditions and defining the question-

naire by nesting an additional canvas within the PocketBee Designer (see Figure 35). Upon zooming into a questionnaire, a similar pipe-and-filter concept is applied, which allows the researcher to drag different questionnaire items onto a canvas and to link them together in the order of appearance on the user interface. Zooming into an item allows definition of this in detail, such as the kind of rating scale used, whether an answer is forced, or whether participants can answer an open question by entering text or by recording a voice message. A branching object supports the branching of questionnaires. In this way, the user can simply define multiple output pipes.

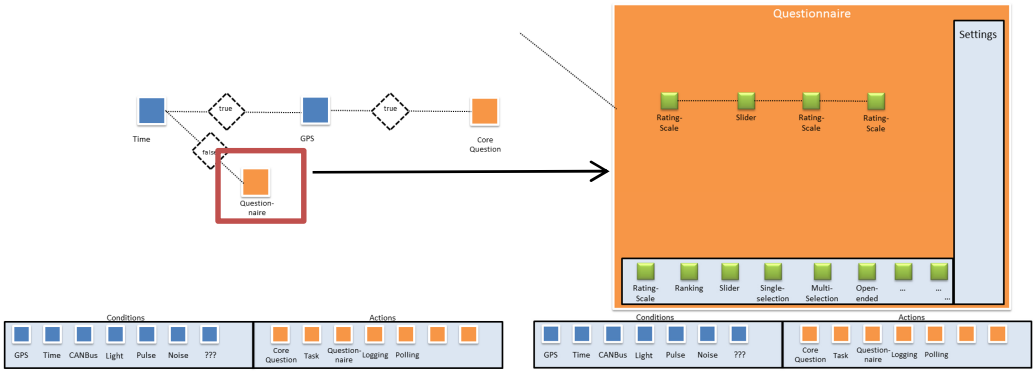


Figure 35: Zoom into questionnaire action opens a new zoomable canvas that allows the placement of questionnaire items in the same style

- *Further actions*

  We also provide a *log* system-action, which allows logging a specific sensor if conditions are met. More complex actions could include the possibility to establish a direct (phone) communication channel to the researcher or further system-actions such as Xensor or MyExperience include: automatically triggering the recording of an external sensor device, such as a camera, for example.

3.4.7.6 The Linking

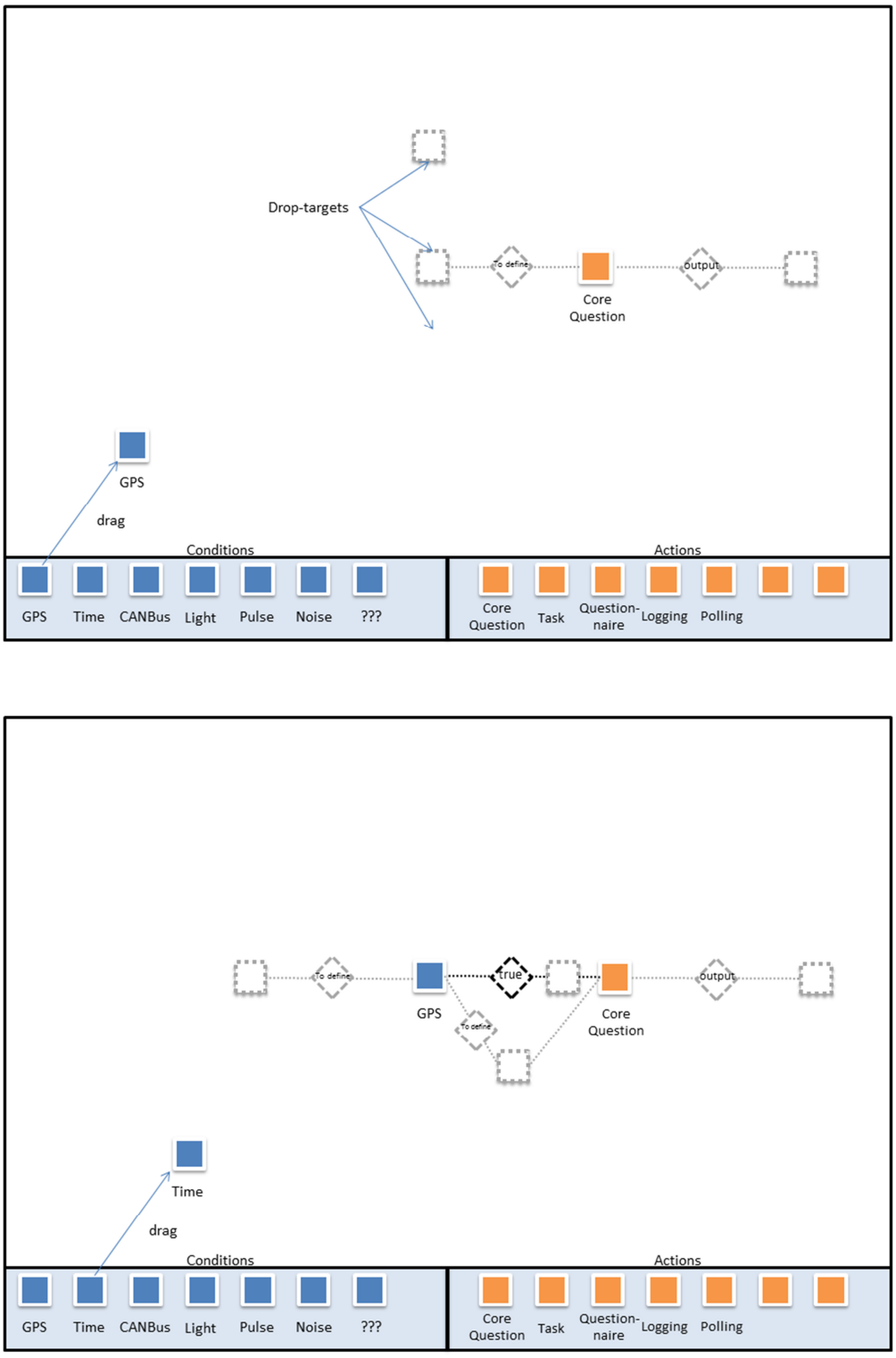


Figure 36: Drop Targets (top) and Boolean connectors (bottom)

Links can be easily established by making use of automatic drop targets that appear as soon as the researcher drags a new object out of the bottom tool bar. Thereby, the researcher can connect conditions by simply dragging and dropping them into a potential target location. This visually resembles the logical layer and provides the primary benefit of the pipe-and-filter metaphor. In this way, it is possible to create complex condition chains without losing the overview. It allows the simple and effective reuse of conditions with different actions or the use of multiple condition chains that end up with the same action. Each condition-object has an input connector and an output connector. A Boolean node allows the definition of the output connector as a true or false output, thereby integrating a Boolean-like AND (true) and NOT (false). Each connector can be used to connect multiple objects, not just one. An action-object also has an input connector and an output connector in case a conditional output value is to be defined. In such a case, a link can be established from the action to a further condition. An additional output selection-object is automatically created between the two, as the researcher needs to define the conditional event for the action (e.g., whenever the participant takes a photograph).

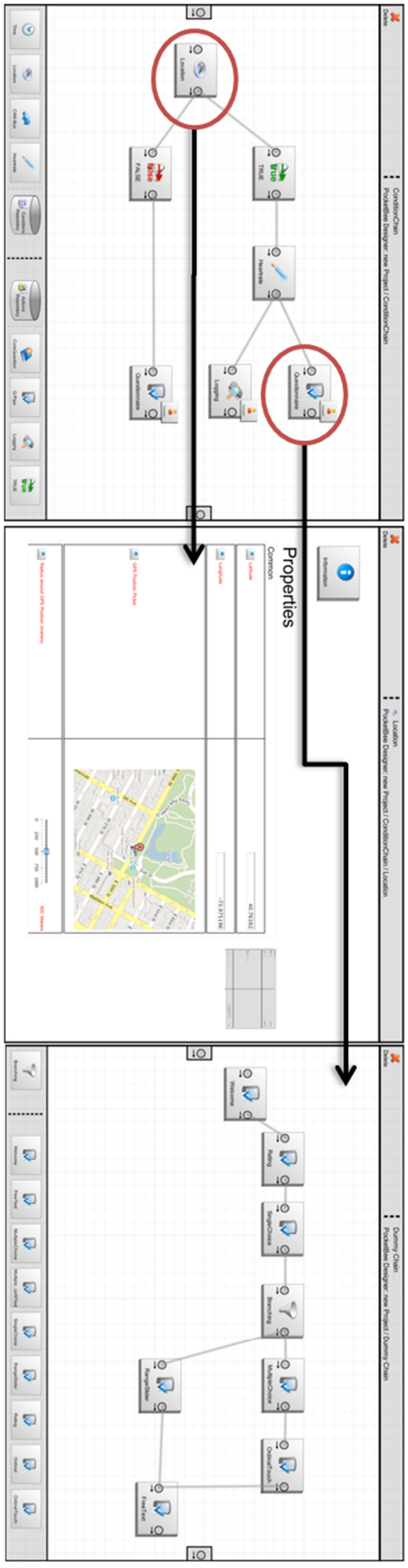


Figure 37: A first running prototype of the PocketBee Designer. Top: Overview of condition-action chain; middle: condition details; bottom: questionnaire configuration.

3.4.8 Conclusions & Future Work

In this chapter focused on PocketBee, we have contributed to the research on diary/ESM tools on various levels. We presented an event-architecture that is capable of supporting the full range of possible research designs (as discussed in Chapter 3.1.4) and allows flexible combination and mixing of designs. This is achieved by building upon the three-level architecture presented by MyExperience and extending it to cover the entire range of possible research designs. The client-side user interface for the participant demonstrates that it is possible to transfer such designs to an easy-to-use and flexible diary user interface, which is not limited to a specific study design. Our case studies showed the utility of mixed study designs as well as the usability of the tool, which was widely accepted and did not cause any drop-outs (within the limited scope of the studies). Furthermore, providing multiple modalities turned out to be beneficial both for the participant in terms of ease and comfort of use and the researcher, who could rely on richer data. We also found that diary/ESM devices are accepted, even if they are provided in addition to the participant's personal mobile phone.

As not only the participant has to interact with an interface in such studies, we presented a conceptual design for a PocketBee Designer interface. It allows transferring the flexible event architecture to complex studies with multiple nested conditions. The concept draws upon the idea of a zoomable user interface in combination with a pipe-and-filter metaphor, which allows the researcher to plug in different conditions with user and system actions and then zoom in to define them in detail. The visualization resembles the modular concept and allows easy integration of new conditions/sensors or actions. Figure 37 shows a first running prototype of the interface.

Compared to existing tools, the main benefits of the PocketBee system are the usability and flexibility of the client-side user interface. Most other tools focus mainly on experience sampling with condition-based designs. They often incorporate typical human-recognition based designs by using workarounds, e.g. by prompting users to capture data in questionnaire items instead of offering a flexible GUI (exceptions, that do have other limitations, such das InfoPal (Jain,

2010) are discussed in 3.4.2). Thereby, they do not include concepts on how to improve the data-gathering process for the participants. PocketBee presents the idea of using core questions to trigger the participants to look out for necessary triggers, a concept that has not been used in any other interface and has proven to be effective in the usability studies. Also the possibility of PocketBee to combine multiple modalities when capturing data is only available to such a degree in InfoPal (which again does not provide advanced condition-based design possibilities).

Besides, the conceptual design for the PocketBee Designer, while not yet evaluated, is the first approach to allow researchers in an easy way the design of complex studies. As the UI concept of a zoomable pipe and filter metaphor has been proven to be very useful and usable in other contexts, e.g (König, Rädle, & Reiterer, 2010) , we do believe that this approach offers lots of potential. From our experience working together with researchers from psychology and social sciences, the need for such a visual design tool is definitely there.

PocketBee offers a variety of aspects for future research. One aspect will be to implement the PocketBee Designer, test it with users and improve it accordingly. In addition, another interesting idea is to increase the support for the analysis process within the Designer. Currently, the researcher must rely on external tools for the most part. However, integrated live visualizations of the incoming data-stream could help tremendously in on-the-fly data analysis. For example, participants that are not responding could be quickly identified and contacted, or the researcher might post additional questions to interesting diary entries.

Completely different usage scenarios of the whole PocketBee system are also tempting, especially in the context of design work as a creativity support tool that allows designers to collect inspiring data along the way and share it with others. Therefore, PocketBee could act as a collaborative design idea capturing tool.

PocketBee would also be well suited to study the diary method itself more in detail and investigate certain effects as habituation, compliance, and panel attrition. For example, due to PocketBee offering a direct link between participant

and researcher, one could easily study how more or less “activity” from the researcher influences these aspects.

It is the eventual goal of the whole project team to make PocketBee available to the public. We have already shared the application with other researchers from sport sciences and medical psychology. An Open Source approach could help us to gather fruitful feedback and might even attract others to actively contribute to the concept and implementation.

3.5 Conclusion

In this chapter, we have presented diaries as a research method for longitudinal field studies in HCI. We have provided a comprehensive overview of research questions and research designs and the relationship of diaries to the experience sampling method (Hektner, Schmidt, & Czikszentmihalyi, 2007). Furthermore, we presented a new classification scheme for diary/ESM studies that unifies these two fields under one umbrella. This classification scheme should inform the design of future diary studies and give researchers an overview of the possibilities they might have in designing such a study.

Examination of applications of the method demonstrates that electronic diaries are becoming more and more predominant. These systems make use of mobile devices (such as PDAs or smart phones) and are therefore inherently mobile. This makes it easier to achieve a closer timing between events and data-gathering, as such devices offer the possibility of instantaneous feedback. Furthermore, these devices incorporate many sensors and functionality that can be useful for diary research, such as integrated cameras or voice recording capabilities. In Chapter 3.4 we presented PocketBee, a multi-modal diary/ESM tool for field research. The system uses a distributed client-server architecture and allows the researcher to design studies spanning the entire range of research designs that have been presented. In addition, the user interface on the mobile device is very easy to use, as several studies have shown, leading to a high degree of acceptance among participants and potentially higher compliance and reduced panel attrition.

Diaries provide a tremendously useful tool for longitudinal research in the field of HCI, and e-diaries in particular push the capabilities and flexibility one step further. They make it much easier to gather longitudinal data in the field over prolonged periods of time and with PocketBee, can be used both for qualitative and quantitative data gathering. Given the increasing size of the displays or tablet variants such as the iPad, one could also include constructive drawing methods such as iScale (Karapanos, Martens, & Hassenzahl, 2009) within a diary tool. Therefore, such tools provide the technological basis for many longitudinal data-gathering methods. To this end, it is especially important that the architecture and interfaces of these tools do not limit the researcher to very specific diary or ESM methods but allow the flexible extension to cover new ideas for data-gathering, such as with PocketBee.

We hope that this chapter encourages other researchers to make use of this approach much more often, and that PocketBee and other diary tools with this combination of functionality and usability become publicly available, so that everyone can incorporate them in their studies.

4 Concept Maps – A Longitudinal Evaluation Method to Assess the Usability and Learnability of APIs

Parts of this chapter were published in (Gerken J. , Jetter, Zöllner, Mader, & Reiterer, 2011).

As discussed in Chapter 2, there are only a few methods that facilitate the tracking of changes in qualitative data over time. In comparison to quantitative data, non-numerical changes are much more subtle. Therefore (as is the case with all qualitative data), much is left to the interpretation of the researcher. However, qualitative data offers tremendous benefits in understanding the WHY and HOW of change processes in detail. This is especially true for situations in which we might not find clear overall patterns of change, but instead highly different change processes for each individual participant. A promising approach by Saldana (Saldaña, 2008) for structuring qualitative data and analyzing it for changes was presented in Chapter 2. However, this approach is limited by the analysis process itself. In the case of qualitative data, we think that a longitudinal research method should address data-gathering as well and should provide techniques that make tracking changes in the data easier.

Considering some of the recent approaches for retrospective designs, such as the iScale method presented in Chapter 2 (Karapanos, Martens, & Hassenzahl, 2009), we find that *constructive techniques* in which participants are asked to create time-dependent artifacts seem especially promising for longitudinal research in general, and not only for retrospective methods. The created artifact makes changes visible, comprehensible, and easier to communicate. Obviously, there is still the need for interpretation (as with quantitative data), but the data basis for this is much less hidden among the otherwise enormous amounts of qualitative data, such as text or video. In this chapter, we explore such a constructive approach in the context of application programming interface usability. Application programming interfaces (APIs) are the interfaces to existing code

structures, such as widgets, frameworks, or toolkits. These interfaces have a great deal of impact on the quality of the resulting system; therefore, ensuring that developers can make the most out of them is an important challenge. However, standard usability evaluation methods used in HCI have limitations in grasping the interaction between developer and the API, as most IDEs (essentially the GUI) represent only part of it. The longitudinal aspect is of special interest here, as an API is not a tool that a developer learns and masters once. Rather, APIs are used when needed and to the extent they are needed. To assess usability, it is therefore critical to analyze the longitudinal perspective, as this will provide an estimate of the API's learnability and also allow insight into individual change processes (i.e., how a user gets past a certain shortcoming of the API or hits a barrier). In this chapter, we present the Concept Map method to study the usability of an API over time. The basic idea of the method is to let users visualize their understanding of an API in a map. A longitudinal panel design is then used to ask them to update this map several times. This allows us to elicit the mental model of a programmer when using an API and thereby identify usability issues and learning barriers and their development over time.

To recapitulate, this research method addresses the needs raised in Chapter 2.2.4.3 for investigation of research questions interested in the WHY and HOW of change processes. It is especially suited for longitudinal panel designs with a small number of participants, i.e., longitudinal case study designs. However, the method is not a "generic" evaluation method (like interviewing); it was specially designed for API usability studies. As such, it should be seen in the context of other constructive approaches, such as iScale (Karapanos, Martens, & Hassenzahl, 2009) or the robot cleaning map previously discussed (Sung, Christensen, & Grinter, 2009). In the following chapter, we will illustrate the method in detail and present a case study to illustrate its benefits and drawbacks. We start with an introduction to API usability and the challenges of conducting evaluation studies in this field.

4.1 Introduction

In modern software development, programming from scratch has become a rare occurrence. This is not only true for updated releases but also for “new” products. Developers often rely instead on existing widgets, frameworks, libraries, or software development toolkits that provide existing code structures for reuse. To access these, application programming interfaces are provided (APIs); although there are many different kinds of APIs, they all serve the same purpose, as Daughtry et al. (Daughtry, Farooq, Myers, & Stylos, 2009) described: “they each provide a programmatic user-interface to a module of code.” As with any kind of interface, some are more usable than others; this can have a tremendous impact on the final product as well as the efficiency of the development process. Advocates for API usability, such as Joshua Bloch from Google, have stressed that

> Good APIs increase the pleasure and productivity of the developers […], the quality of the software they produce, and ultimately, the corporate bottom line. Conversely, poorly written APIs […] have been known to harm the bottom line to the point of bankruptcy. *(Bloch, 2005)*

A number of researchers have begun to investigate the usability of APIs in more detail in recent years, with McLellan et al. (McLellan, Roesler, Tempest, & Spinuzzi, 1998) often cited as having conducted the first formal usability study of an API. Since 1998, there have been quite a few studies on different design aspects, such as the use of different patterns (e.g., (Ellis, Stylos, & Myers, 2007)) or API documentation. In addition, several books and papers providing API design guidelines have been published (Cwalina & Abrams) (Tulach, 2008). At the CHI 2009 conference, a special interest group (SIG) took place on API Usability (Daughtry, Stylos, Farooq, & Myers, 2009) to discuss the challenges of designing a usable API. As one outcome, the organizers have created a website that includes a collection of useful resources and links to papers on the topic.

Figure 38: A concept map of the ZOIL API

One area within this field with limited prior research is the data-gathering methodology used to assess the usability of an API. Most methods have essentially been adaptations of existing HCI usability evaluation techniques, such as usability tests and inspection methods. Since APIs are fundamentally different from the graphical user interfaces for which these methods were designed, we believe that there is a need for evaluation methods that have been specifically designed to address the specific nature of an API. Because the GUI (which allows researchers to directly observe the interaction with an interface) is missing, direct observation methods are more vulnerable to subjective interpretation. Inspection methods require a high level of knowledge about the API and API programming in general by the analyst. In addition, writing a piece of code is often a tedious process lasting days if not weeks; thus, depending on the observational approach and the complexity of the API, it can be difficult to define ecologically valid tasks that would fit into a 1-2 hour observation session. Further-

more, using an API is a constant learning process: developers seldom read documentation in advance, instead searching for examples or documentation as they work. Therefore, a research method for API usability should be able to measure this learning process over time and allow the researcher to identify learning barriers.

The evaluation approach that we will present in this chapter will address and contribute to this issue. It is based on the concept-mapping technique (see Figure 38) familiar from learning theories (Novak & Gwon, 1984), and allows the researcher to elicit the developer's mental model when working with an API by making the interaction visible. Furthermore, it is especially useful in a longitudinal design, as our method is designed to permit the tracking of changes within qualitative data over time - a common and difficult to address challenge in longitudinal research, as we (in Chapter 2) and others (Courage, Jain, & Rosenbaum, 2009) have discussed. Our method can be used to assess the learning barriers that developers encounter when working with an unfamiliar API as well as their evolution over time. In addition, our method is easy to apply in practice, because it uses hands-on materials and may include a wide variety of possible metrics. In the following sections, we will first review existing literature of API evaluation methods to discuss the challenges that a method should address. We will then present the method in detail, outlining the materials, the design rationales, and the data-gathering process. Finally, we will discuss the application and analysis possibilities of the method by presenting a case study of an API evaluation with university students who were given the task to create a software prototype with the help of an unfamiliar API.

4.2 Challenges for the Evaluation of an API

In reviewing the literature on API evaluation methods, only a few papers focus specifically on the data-gathering method (e.g., (Farooq & Zirkler, 2010) (Clarke, 2004) (McLellan, Roesler, Tempest, & Spinuzzi, 1998)). However, there are quite a few papers that present, discuss, and evaluate certain design

choices, such as the specific patterns of an API. We can identify three different principal purposes for studying the usability of an API. The first is to support the development process of an API, following the user-centered tradition of usability engineering lifecycles. In this case, studying the usability has the goal of obtaining answers to questions such as how easy it is to learn the API, how efficiently can it be used for specified tasks, or which areas are difficult to use and lead to misconceptions in the programmer's understanding. The second purpose is to derive design principles and a theoretical foundation for the design of new APIs, as discussed in the introduction: studying and analyzing existing APIs serves the purpose of understanding how people actually use them, which can then help us to design better APIs in the future. The third purpose would be to conduct comparative studies of APIs. This is especially important when companies have to decide which of several competing APIs they should introduce in their software development process, but also for marketing purposes when launching a new API.

4.2.1 Data-gathering

A tremendous challenge for the evaluation of an API is that using and interacting with an API is much more subtle than using a standard software application and therefore more difficult to observe and analyze. Similar to the use of command languages, we can normally only analyze the end product of complex thinking processes within the user. Accordingly, definition of mistakes or errors during the observation of users is not necessarily straightforward, since there are many ways to reach a goal.

The most common approaches to studying the usability of an API have been lab-based usability tests in combination with the thinking-aloud protocol. In the previously cited study by McLellan et al. (McLellan, Roesler, Tempest, & Spinuzzi, 1998), four programmers from an API target group were given the task to analyze and understand a code example that used calls from the API. The participants were asked to think aloud while trying to understand the code and to express what information about the API they would need to reproduce

such a code sample. They were also asked what additional features they would expect from the API from what they had seen, allowing the researchers to assess how users of this API might perceive its limitations (Myers, Hudson, & Pausch, 2000). As the participants were allowed to ask questions of an API expert, one could describe this approach as a kind of co-design for API development. In Klemmer et al. (Klemmer, Lie, Lin, & Landay, 2004), the authors conducted a more traditional usability test, with seven participants using the *Papier-Mâché* toolkit for developing tangible user interfaces. Participants were first introduced to the toolkit and then were asked to complete three typical programming tasks using it. Thinking aloud as well as participants' Java code was then used to analyze the usability of the toolkit. In a similar way, Heer et al. (Heer, Card, & Landay, 2005) analyzed the usability of their *prefuse* toolkit. An interesting variation of this approach was proposed by Beaton et al. (Beaton, Myers, Stylos, Jeong, & Xie, 2008). In their approach, participants would first write in pseudo-code what they would expect in the API for a certain task and would then perform the real task using the API. Thereby, the authors suggest, one could better assess the mapping between the user's mental model and its real-world counterpart. All these approaches had the primary goal of finding usability flaws within a specific API rather than generating knowledge for a theoretical basis for API design. Contrarily, de Souza et al. (de Souza, Redmiles, Cheng, Millen, & Patterson, 2004) performed an extensive field study to understand how APIs are used in practice, which roles they serve, and whether their use has only beneficial purposes or also drawbacks. The authors spent 11 weeks on-site at a software company, conducting non-participant observations and semi-structured interviews; they also had access to documents about the processes and to discussion databases. In a grounded theory approach, their data was analyzed and continuously enriched with new observations and interviews. The nature of such a study obviously makes it inappropriate for analyzing the usability of an API during the development process; however, more focused, short-term field observations could help to define requirements for a forthcoming updated version of an existing API, for example.

In addition to these methods using direct involvement of end users (programmers), there has also been some research regarding analytical inspection methods, comparable to usability inspection methods such as cognitive walkthroughs or heuristic evaluation. Obviously, the main advantage here is that no real users are needed; this may facilitate testing, since the target group of an API often is spread around the world and is not as easy to get into a lab as the potential iPod user. Farooq and Zirkler (Farooq & Zirkler, 2010) presented a method called API Peer reviews, which is based on cognitive walkthroughs adapted to APIs. The approach has been used within Microsoft in addition to usability tests. It is a group-based usability inspection in which different members of the API development team serve different roles: for example, the feature owner is the one whose part of the API is under review, and some of the team members serve as reviewers. During a 90 minute meeting, the goal is to walk through a specific part of an API while attempting to recreate a typical scenario of use. The reviewers comment on this by trying to put themselves in the role of users. The method has proved to be highly scalable and to have a very good benefit-to-cost ratio. Nevertheless, the authors see it as an addition to usability testing rather than a replacement.

4.2.2 Metrics

The studies cited above have used both qualitative and quantitative approaches to assess usability. Purely quantitative measurements include task-completion times (Ballagas, Memon, Reiners, & Borchers, 2007) (Ellis, Stylos, & Myers, 2007), lines of code (Klemmer, Lie, Lin, & Landay, 2004), and number of iteration steps needed (Ballagas, Memon, Reiners, & Borchers, 2007). While these can help in the comparison of different APIs (Ellis, Stylos, & Myers, 2007) they can only indicate usability issues in a rather broad sense. More detailed qualitative analysis of the think-aloud protocol and video observation data can assist in identification of deeper usability issues. Here, the work of Clarke (Clarke, 2004) has been rather influential. Clarke used the cognitive dimensions framework (Green & Petre, 1996) and adapted it to fit the requirements of API usability

evaluation. By using this framework, researchers can cluster findings into different categories (e.g., API Viscosity or Consistency) and thereby identify which higher-level concept of the API might be problematic. Farooq and Zirkler also relied on this framework to cluster the findings of their API Peer Review approach (Farooq & Zirkler, 2010).

Ko et al. (Ko, Myers, & Aung, 2004), on the other hand, identified six learning barriers of an API (e.g., selection barriers or information barriers) in a large field study which can again be used to cluster qualitative data. Identifying such learning barriers is one step in assessing the threshold of an API – basically, how difficult it is to achieve certain outcomes with the API.

Myers et al. introduced the *threshold* and *ceiling* concept as quality criteria: “The threshold is how difficult it is to learn how to use the system and the ceiling is how much can be done using the system” (Myers, Hudson, & Pausch, 2000). In most of the studies cited so far, the goal was to identify the threshold or the barriers within the API that seem to increase the threshold. The ceiling, on the other hand, defines what is achievable with an API. Instead of looking at the process, one can look at the artifacts that can be created using a specific API and thereby determine its value and quality. Common approaches here are case studies that display a wide range of possible systems (Heer, Card, & Landay, 2005) (Klemmer, Lie, Lin, & Landay, 2004).

In summary, the most common data-gathering approaches are usability tests, thinking aloud, inspection methods, and in some cases field observations. From an analysis perspective, the metrics include straightforward aspects (such as task-completion time and lines of code) as well as more theoretically grounded analysis frameworks (such as the cognitive dimensions).

We find that the current approaches seem insufficient to address two major aspects: first, in the case of observation or inspection approaches, most studies are limited to one or perhaps a few hours. As a result, the tasks involved are rather simple and generally “pre-defined” with given code samples. More complex or real-world tasks in which developers can use the API for real projects

are rare and difficult to integrate into study designs, although such tasks would provide very valuable input regarding the usability of an API in real-world situations. Second, it is difficult to assess eventual changes in learning barriers or the threshold of an API during a single session. One could assume that barriers shift during longer usage durations and that thresholds might be perceived differently after some time. Both of these aspects can be addressed by using a longitudinal study design that gathers data at more than one point in time (Taris, 2000). What is still needed is an appropriate data-gathering method that would enable integration of more complex tasks and observation of these changes. Most approaches rely on direct observation or inspection; however, given the task of coding a piece of software, we can see a certain value in retrospective approaches that might allow users to better reflect on the pros and cons of an API. Simple retrospective interviews seem insufficient for this purpose, as they would lack a proper artifact to trigger the discussion with the participant. In the following section, we will present the Concept Map method, which incorporates a longitudinal panel study design and a visual representation of the API usage and therefore directly addresses these issues.

4.3 The Concept Map Method

Novak (Novak & Gwon, 1984) introduced concept mapping in the late 1960s and early 1970s as a research method during a longitudinal 12-year research project that assessed how children's understanding of science concepts changed over time. Concept maps can be described as visual knowledge representations with nodes and edges. Each node represents a concept and is linked with one or several other nodes via edges. The edges are typically directed and labeled to describe the nature of the connection between the two nodes. Originally, the device was defined as a top-down diagram to decompose hierarchical relationships within a main concept. However, it has since been applied in a number of variations, including non-hierarchical, flat structures. Novak originally introduced the method to improve biology teaching; it has subsequently been shown to have great value in student learning for a variety of top-

ics and teaching situations (Eppler, 2006), both as a learning strategy and as an instructional strategy. It has for example been applied as a means of assessing students' understanding of science concepts (McClure, Sonak, & Suen, 1999). In HCI, concept maps have been implemented as creativity and structuring tools similar to mind maps: for example, during the requirements phase in a usability engineering process (Barksdale & McCrickard, 2010).

Given the nature of APIs, we propose concept maps as an evaluation and assessment method to elicit the programmer's mental model of an API. Thereby, we will be able to identify misconceptions and problematic areas and assess how these change over time.

4.3.1 Main Idea

An API, by definition, is an interface between two distinct pieces of software code: one is the application that is under development, and the other is a more general framework or SDK for which the API provides the interface. Our concept-mapping approach asks participants to visualize this relationship between their own piece of code (which can be a given task or a real application) and the API. This takes place during a 30-60 minute observation session, which is videotaped and includes a thinking-aloud protocol. For each participant, this session is repeated (e.g., once a week over a five week period), depending on the complexity of the API and the application. During the later sessions, the users do not start from scratch; instead, they are handed their concept map from the previous session and asked to change everything that they no longer perceive as being a correct representation of their mental model. This is an important aspect, as we do not inquire how their understanding of the API has changed (which would be much more difficult to answer) but rather ask them to update their own artifact. How their understanding has changed is then implicitly reflected in the changes they make to the map. Analysis of these maps with the help of API experts makes it possible to understand misconceptions of or usability problems with the API. Given the graph-based structure of such a map, we

are furthermore able to (digitally) compare it with a “master” map created by the API’s developers or API experts.

4.3.2 Design Rationale and Materials

The method is designed with hands-on materials, making it easy to implement in any environment. In the following section, we will present the materials required and discuss the design rationale and possible design choices behind them. The method was developed over the course of several case studies and refined in many aspects along the way. In some cases, we will offer different design possibilities depending on the goals of the study.

- *Participants:* Before we discuss the method specifics, we would like to address the issue of participants. As we are studying API usability, not every person qualifies as a participant, since specific skills are required to be able to use an API. First, participants should have at least 2-3 years of programming experience – otherwise, most issues that would be revealed during the study would not be API specific but rather general programming issues of the participants. It is furthermore important that this experience is based on the same or very similar programming languages, to avoid different mental models of the language being a factor. In the best-case scenario, the participants would be actual users of the API (if it is already deployed) or would be expected to become users in the future. The motivation to take part in such a study might thereby be increased as well, as the study would offer a learning experience that participants could benefit from. In our case studies, we selected students in Computer Science who were expected to work with the API during a lecture and afterwards in future projects.
  Another issue is the number of participants. The Concept Map method is an in-depth method that requires time and resources to analyze the material in detail. We had good experiences in our case studies using 10-12 participants working in pairs. Should there be more participants, we would advise scheduling them with a chronological offset to reduce the complexity of the study. As mentioned, we paired our participants. This is not a necessity, but

provides advantages. The Concept Map sessions, as we will illustrate, produce an enormous amount of context information. Asking a pair of participants to create a shared concept map essentially requires them to discuss aloud and negotiate every decision with each other. As a result, the video material from such a session includes not only the maps but also the reasoning during the creation process. While we have not tested this, we assume that asking a single participant to think aloud during a session would not reach the same level of depth of reasoning as in our experiences with think-aloud protocols in usability tests. However, some drawbacks must be considered as well: certain problems could stay hidden, since it would be sufficient if only one participant had the correct understanding. Similarly, the mapping session tends to be more a constructive act than pure elicitation, as the two participants have to find a common language and understanding.

- *The "mapping" session:* In the case-study section, we will illustrate in more detail what the mapping session looks like exactly. In brief, participants create a map (starting from scratch) that shows the relationship between the API and the prototype/system they are working on. Asking them just to visualize the API would be a very artificial task, but asking about the interrelationship between the API and the system under development requires participants to adapt their thinking processes while programming and using the API. We will provide the set-up for creating the map in this section. Important to note is that a researcher and API expert should be present during the sessions. The researcher's job is to oversee the API mapping process and to instruct participants in what kind of symbol language to use, when to rate concepts, etc. The API expert should be on-site because during the process participants often recognize or remember problems they have had with the API. An API expert can note these accordingly, facilitating the data-analysis process by commenting on them. This may be important, as participants' progress may otherwise be hindered. In longitudinal designs, the researcher should avoid a situation in which a participant is "stuck" for long periods, as attrition is nearly inevitable in such a case. In particular for studies analyzing

learnability, it is sensible to implement such feedback protocols, as it does not help the researcher to only know the first of possibly several learning barriers one could come across at later stages. See, for example, (Grossman, Fitzmaurice, & Attar, 2009) who implemented a Question-Suggestion protocol that we adapted in one of our case studies (Gerken, Jetter, & Reiterer, 2010a).

- *Data-gathering waves and schedule:* In a longitudinal approach, the question arises of how much time is necessary for the study to be effective. As we have made clear in Chapter 2, the overall duration is not necessarily a factor in longitudinal studies, but rather the important aspect might be the number of data-gathering waves and the expected change processes that one seeks to uncover. The number of repeated sessions required strongly depends on the complexity of the API, the nature of the task, and the experience of the users. A more complex API or task, or less experienced users will automatically result in more sessions required to achieve leveling-out in the maps (i.e., no more changes in the data). In our studies, we used at least four iterations to be able to measure changes as well as a level of stabilization. We used weekly schedules because our participants were students, each with their own various obligations. In the case of professional programmers or participants who can devote all their time to the study (e.g., because they are working on a real product with the API anyway), the time in between sessions could be much shorter, probably 2-3 days.

- *A modified corkboard/whiteboard:* We have implemented the method both on a table and on a vertical corkboard. While the table allows more people to position themselves around the map, the vertical board has the advantage that it allows the user to step back and gain an overview, which we consider to be an essential advantage of that setting.

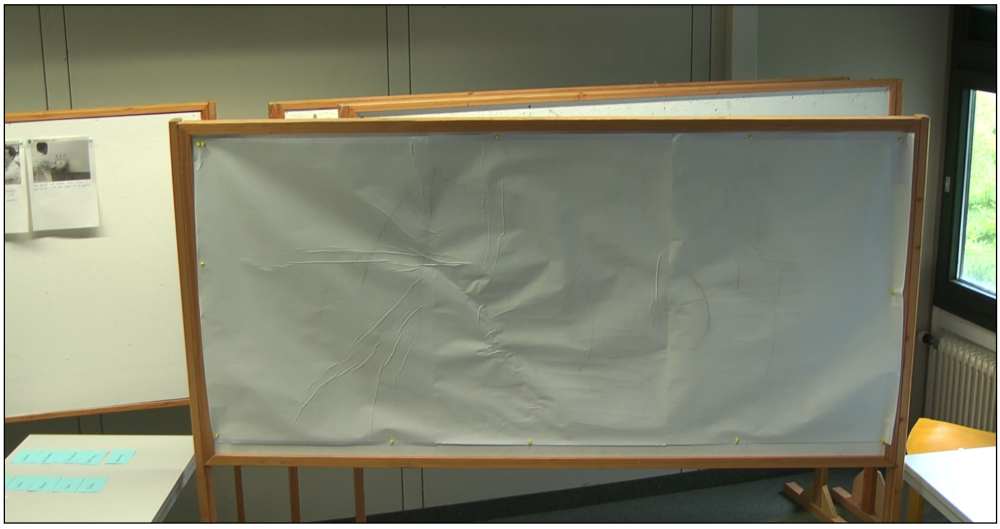


Figure 39: A “modified” vertical pin board

As we want to allow participants to easily place concepts on the map as well as to change the placement and any links they have created, a huge whiteboard would be the best solution. A hands-on alternative that we used during our second study is a modified pin-board with painter foil covering it (see Figure 23). This allows participants to pin concepts to the board (as on a pinboard) and also to draw and remove connections (as on a whiteboard)..

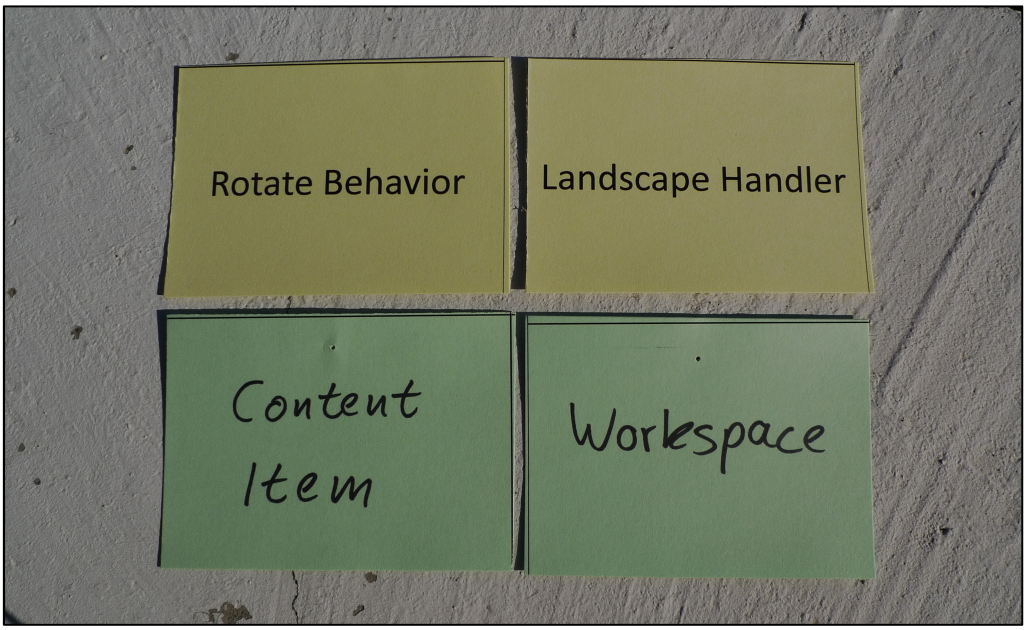


Figure 40: Yellow API concepts and green prototype concepts

The concepts: In our studies, we used 7.5x10.5cm cards for each concept (see Figure 40). Depending on the goal of the study, it is possible either to pre-define concepts or to permit participants to define concepts themselves. A more explorative study would prefer the latter, while a more controlled setting with specific parts of an API under investigation should pre-define concepts. This allows easier comparison of concept maps between users or with a master map, enabling quantitative data analysis. What is a concept? The granularity of a concept can be adapted to the research goal as well. A concept can be a certain method, a class name, or a higher-level construct that includes multiple classes. It can also be detached from the actual code by using an abstract or a user-centered perspective. For example, if the API is responsible for handling the input modalities, one concept could be “Input modality”, or this could be broken down into “mouse input”, “touch input”, “voice input”, etc. By using different levels of granularity for different parts of the API, the researcher can define which aspect is under close investigation (the detailed part) and still assess the overall understanding of the entire API. We further distinguish between API concepts and what we call “prototype concepts”, which include the concepts for the piece of software the participant is writing. The task for the participant during the concept-mapping session is to connect the prototype concepts with the API concepts by drawing a line and adding a label to it that further explains the connection. Basically, we ask the users to visualize the processes between the software and the API.

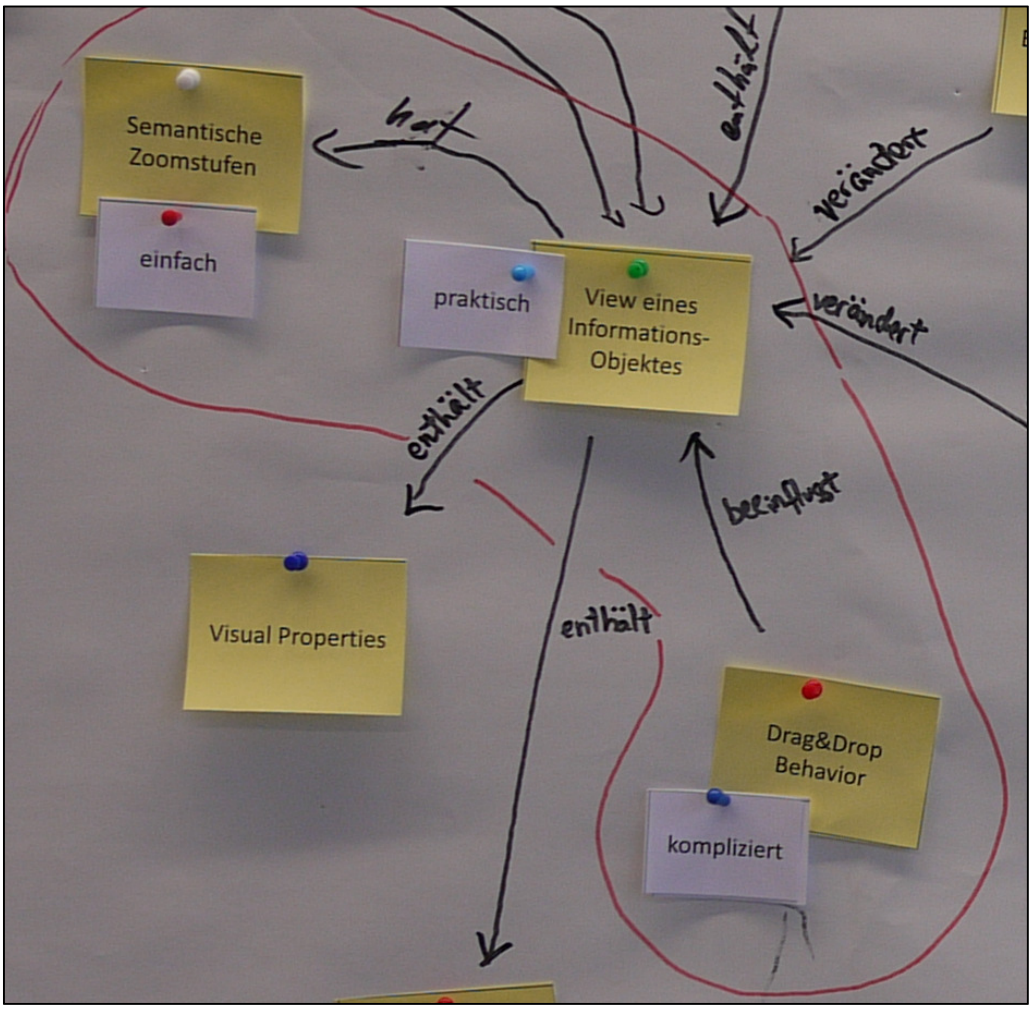
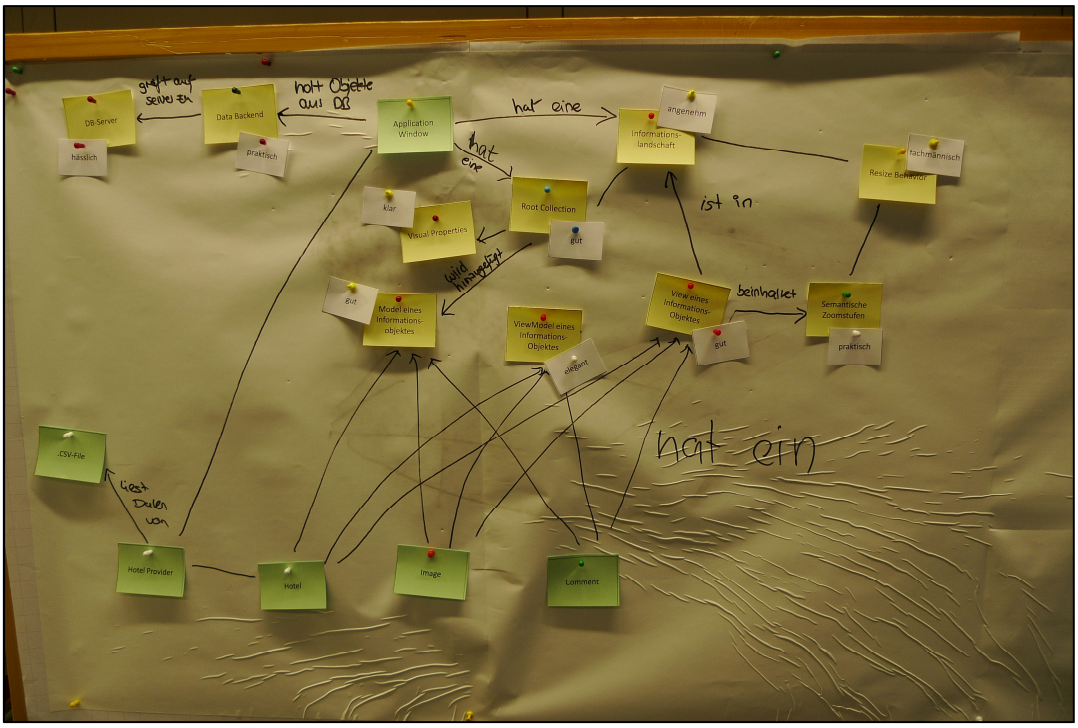


Figure 41: Adjectives (e.g., easy, practical) attached to Concepts (semantic zoom level, view of information object) and a problem area

- *Rating concepts and indicating problem areas:* The method includes two additional tools to help identify and understand potential usability issues (see Figure 41). First, the participants are asked to assign one of several pre-defined adjectives, which are also written on individual cards, to each concept at the end of a session. These adjectives are presented as contrasting pairs of adjectives as in a semantic differential. We have used a set of eight pairs, including convenient – inconvenient, easy – complicated, and beautiful – ugly. Participants are only allowed to assign one adjective per concept; they are to choose the word that best expresses their feeling. The main idea here is to quickly identify the concepts that trigger a positive feeling and

those that trigger a negative feeling. Utilizing the adjectives allowed us to use the same approach as with the concept map cards; however, other emotion or mood measures could be applied here as well. The second tool is asking participants to indicate problem areas by drawing a red line around those concepts with which they have had the most trouble when they were using the API. In this way, individual concepts as well as a whole group of concepts can be marked. We have found that asking participants to indicate these areas quickly triggers responses explaining the problems and thereby provides tremendous help in understanding the relevant usability issues. Again, we think that the presence of the artifact helps participants to talk about such issues more easily, as they can visualize and assign the problem to a concrete object.



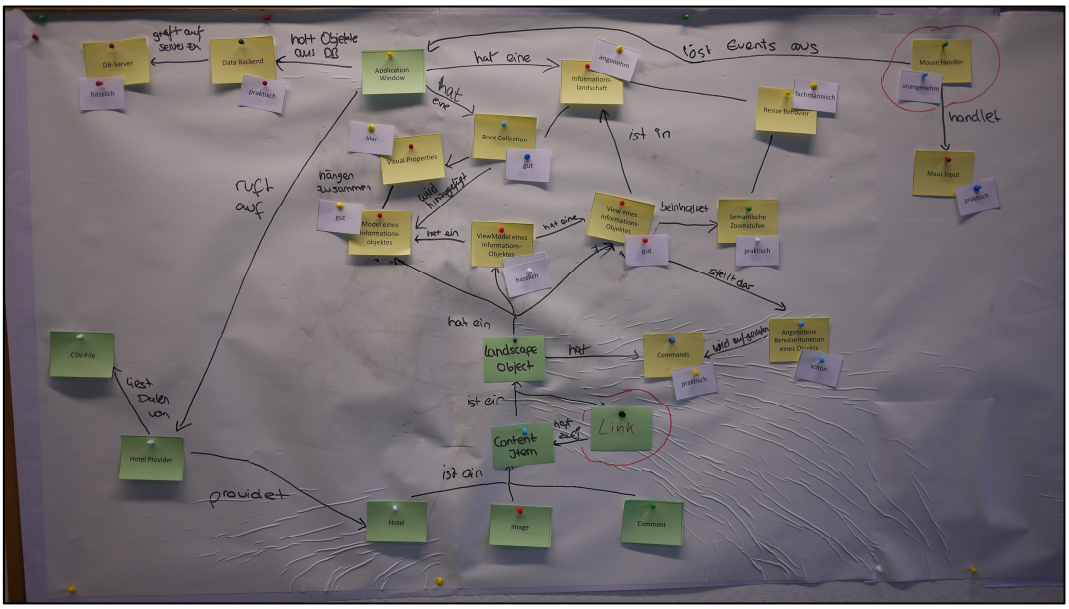


Figure 42: Concept Map session 1 and 2 from group 2

- *Longitudinal Panel Design:* As discussed in the previous section, one principal goal of the method is to make changes over time visually comprehensible. This means that the method can be most effective when there is time for a data-gathering design that includes multiple sessions that build on each other. The idea here is that participants will continue to work on and refine their concept map during each session, given that they have continued to use the API either with predefined tasks or for their real work (see Figure 42). First, they are asked to review all the concepts and connections and are encouraged to think about currently unlabeled links and the map structure. It is thus important to provide a flexible map background, such as a white-board or the modified pin-board. Otherwise, changes would be tedious and participants would be reluctant to make alterations. Changes always hint at an altered or extended understanding of the API and thereby indicating potential problem areas as well as the type of false positives one may come across in a usability test: some aspects of an API might just require some time to learn. This “change and update” procedure is also used for the adjective ratings and the problem areas. Regarding the former, changes are effected by placing a new adjective on top of the old one, making it easy to re-

capture the process in the end. For the analysis, the most interesting points are when participants change from a negative to a positive adjective or *vice versa*, indicating a clear change in perception of this specific concept. Problem areas can be removed, reduced in size, or enlarged; users simply erase the drawing or change it accordingly. This gives researchers an understanding of the complexity of a problem, which is additionally supported by the thinking aloud. Again, being asked to make such changes often triggers users to explain them.

In addition to the clear advantages of the longitudinal design, the method can also provide valuable input in cross-sectional designs – for example, as an addition to a usability test. In this way, one could assess knowledge of an API prior to and after the test. Having this externalization of a user's mental model can furthermore enhance interviews with experienced developers – not to test their understanding, but to understand their knowledge.

4.4 Case Study

The Concept Map method has been developed in an iterative process that included two case studies, which were used to test different variations of the method (e.g., table or vertical board, pre-defined or user-defined concepts). We used a framework for building zoomable user interfaces that was under development in our group as a testbed during the studies. In this section, we present our second case study in detail (the first is documented here (Gerken, Jetter, & Reiterer, 2010a)). The purpose of this section is to illustrate a subset of our study results as empirical evidence of the usefulness of the method, as well as to provide more specifics about the possibilities for data analysis.

4.4.1 The ZOIL API[10]

The **Z**oomable **O**bject-Oriented **I**nformation **L**andscape (ZOIL) API provides access to the ZOIL framework, which is deployed as a software framework written in C#/XAML for .NET and Windows Presentation Foundation (WPF). It provides programmers with an extensible collection of classes covering a wide range of functionality (e.g., ZUIs, client-server persistency, and input device abstraction). Generally speaking, it serves as a toolkit for developing zoomable user interfaces in the context of reality-based interaction and Surface Computing (Jetter H.-C. , Gerken, Zöllner, & Reiterer, 2010). During the study, both the framework and the API were still under development and were not “finished” products.

4.4.2 Study Design and Procedure

We conducted this study as part of a course about visual information seeking systems. The computer science students were given the task to create a prototype of such a system using the ZOIL framework, which they had never used or

[10] The ZOIL framwork is available as open source: http://zoil.codeplex.com/

seen before. However, they were familiar with the C# language. Eleven students participated, split into five groups of two users (in one case, three). This allowed us to implement “discussing aloud” as a variation of thinking aloud during the concept map sessions for a better understanding of the users’ reasoning. We applied a longitudinal design with five sessions over five weeks (one session per week), of which the first was an introduction session. During the other four meetings, the participants were asked to create and modify their individual concept map. Each session lasted about 30 minutes. The overall programming task was split into four milestones; after each session, the milestone for the next week was handed out to the students. In this way, we could recreate a realistic setting in which the task would require users to gain a deeper understanding of the API as time passed.

Concepts: We created a master map of the ZOIL API prior to the study, which took two API developers about three hours. Based on this master map, we preselected 24 concepts. These focused on three aspects of the API/framework: the input handling, the MVVM (Model-View-ViewModel) pattern required to create objects in the zoomable canvas (the application window), and the attached behavior pattern, which allows users of the API to easily attach functionality to any object without having the object to implement it in its class hierarchy. Participants were not allowed to add concepts, as we wanted to control this variable for comparison between groups and the master map. We also provided “prototype” concepts that users were allowed to extend during the sessions in order to reflect their specific implementation of the given task. All API concepts were handed out to the participants in the first session, and they were advised to use these concepts in the map, referring to them in any way. As students were learning the API and the framework during the task, we expected their understanding to change over time, which would then be reflected in their use of concepts on the map.

Procedure: The first session was used to present the programming task and explain the concept map approach. We did so by asking users to build a concept map of the “driver-car” interaction, with the car representing the API and

the driver representing the prototype. In the second session, users had worked with the API for one week and were asked to create a first concept map. We presented the materials, including the modified pin-board, different markers, the API concepts, the prototype concepts, and the adjectives. Usually participants started by flipping through the available concepts and using a table to get an overview. They then began to pin the known concepts onto the board and to connect them with links. They were asked to discuss their decisions with their teammate but were advised that the researchers would not interfere with their task. After on average 20 minutes, participants indicated that they had finished their map. They were asked to once again review the map and check the connections and labels. Finally, we asked them to assign adjectives to the API concepts and to mark any problem areas by drawing a red circle around the concepts; the next milestone for their programming task was then presented. In the subsequent sessions, participants were first asked to review their existing map and change anything that they would now consider to be an incorrect representation of their mental map. The next step required them to extend the map, reflecting the programming done during the week, and to add any additional concepts they had come across. They again revisited the adjectives and the problem areas and made changes accordingly. Each session was videotaped and still photographs were taken of each concept map at the end of the sessions.

4.4.3 Data Analysis

In our understanding, a useful evaluation method must be flexible both in terms of how it can be applied and in terms of the possible measures that can be derived from it. The Concept Maps method provides a large variety of possibilities for data analysis. In this section, we will illustrate the different steps needed and exemplify these with results from our case study.

Step 1 – Digitizing the map:

The method is designed in such a way that the resulting maps can be represented in the GraphML standard (http://graphml.graphdrawing.org/) by using a

graph editor such as yEd (http://www.yworks.com), with concepts and adjectives represented as nodes and problem areas as groupings. In our case study, we reproduced one map from each session (the “final” map, example in Figure 43), resulting in 4 maps per group and 20 maps total (duration for this step: 4 hours). It can also be interesting to include intermediate maps from within sessions, if one is interested in this level of detail.

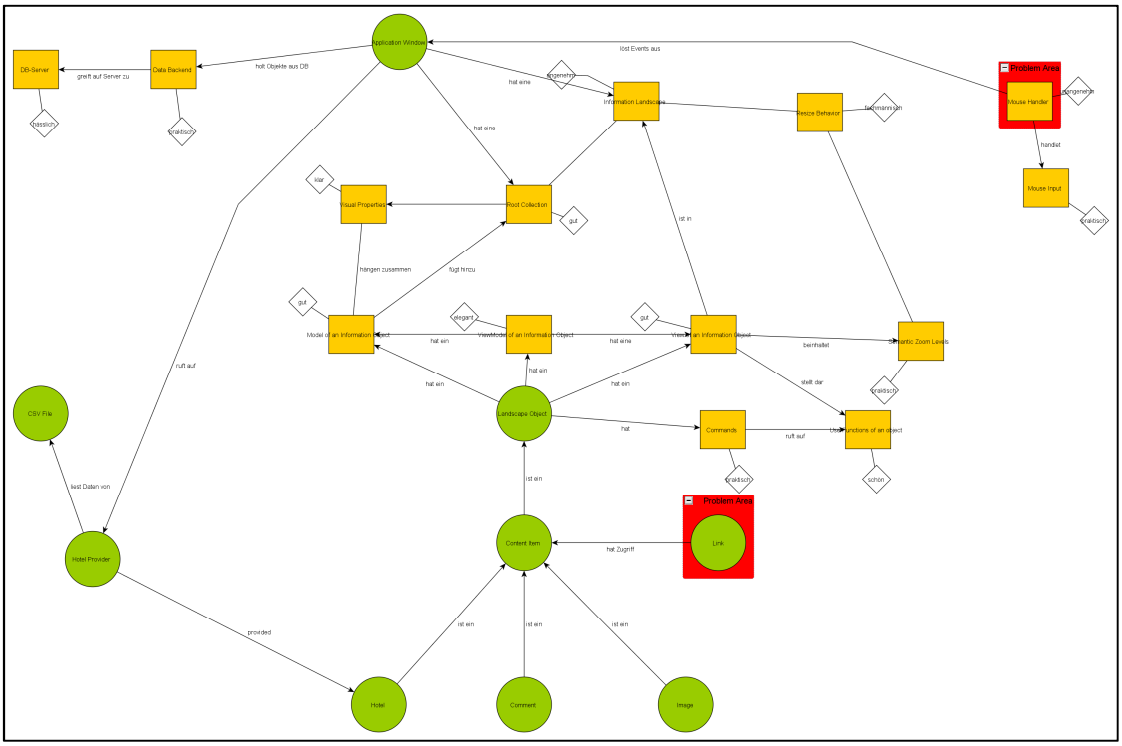


Figure 43: Digitized map of group 2, session 2 (compare to Figure 42 for the still image)

Having this digital representation helps the analyst to identify interesting sections, since the visual noise of a still photograph is cleared away and aspects such as the problem areas stand out more clearly. Furthermore, we can use additional tools for graph analysis, which we will show in Step 3.

Step 2 – General analysis of concepts, adjectives, and problem areas:

Before analyzing individual maps in detail, a more quantitative and general approach can be helpful in identifying potential usability problems and misconceptions. At first, we can check which concepts have been added to the map during

which session. We can cross-check this with the milestones for each session. If a concept were added to the map even though the part of the API it represents had not been used at this point, this could indicate that users were able to anticipate parts of the API that they had not used before. However, if a concept is missing even though the milestone clearly asked the participants to make use of this specific API part, this could indicate that they did not use or understand a necessary part of the API. We can also compare the use of concepts across participant groups and identify how similar the groups are to each other and whether there are similarities regarding the use or neglect of concepts. In the case study, the concept “Landscape Handler” is easily identified as a problematic candidate. This concept refers to the part of the API that captures input events from different input modalities and forwards them to the zoomable canvas (which acts as a view). By comparing the groups, we can see that only two of our five groups integrated this concept into their map, both during the second session. This is correct, since the milestone for this second session was to integrate mouse input into the prototype. However, all of the other groups are missing this concept. Examining the maps of the two groups who made use of the Landscape Handler, we can furthermore see that only one group used it correctly. The other groups connected the Mouse Handler concept directly to the view. While this understanding still resulted in a working prototype (most probably by copying code), the concept maps reveal that these users did not understand the abstraction layer that the landscape handler introduces. The integration of additional input modalities would therefore cause problems and require more time. Therefore, we can clearly state that this part of the API lacks clarity and should either be refined or better documented.

A major benefit of the Concept Maps method in comparison to existing approaches is its ability to capture the dynamics of use, which also refers to the learning of the API and helps prevention of “false positives”. For example, looking at the adjective ratings from this perspective can be very helpful. We can use a simple Excel table to visualize which adjectives have been assigned to which concept at what point in time and whether this changes at some point. Table 1 illustrates this for one of our groups in the case study (group 1). We can

easily see that the *View_Information_Object* and the *ViewModel* of the MVVM pattern were assigned a negative adjective during the first session, which was later changed to a positive adjective (accompanying corrected links between concepts), indicating the overcoming of a learning barrier. Other groups displayed similar behavior; however, in some cases, the negative adjective remains. In such cases, the knowledge of an API designer can be very helpful to resolve conflicts between functional and non-functional requirements (the utility of the MVVM pattern vs. the learning issues). In this table, we also show whether a concept was part of a problem area or not (the red frame around adjectives). The *DB Server* concept was assigned the adjective “complicated” during the first three sessions and “confusing” during the last session. It was furthermore marked as being part of a problem area during the second and third session, but not in the fourth. We interpret the choice of adjectives and assignment to a problem area to mean that the users found a way to get the *DB Server* to work, but even in the end were not quite sure how they had managed it. Thus, the negative adjective remained, but the problem area disappeared. In this example, analyzing the final (functional) code could lead to the wrong impression that the API was well understood (a “false negative”). All in all, we think that concept maps allow a more objective measure of understanding by investigating the dynamics of the learning process.

We can also confirm here the aforementioned issues with the input handler concepts, such as the Landscape Handler and the Mouse Handler. Only one group did not assign a negative adjective to either of the two at some point. The other groups also frequently assigned problem areas to this part of the API (as in Table 9), again indicating some clear misconceptions and usability issues.

Table 9: Adjectives assigned to concepts over time. Each column represents one session and each row one concept. Black = concept not yet added to the map, empty: concept added, but no adjective assigned.. green: positive adjective, red: negative adjective. Red border: part of problem area.

Group 1				
Concepts/Session	G1S1	G1S2	G1S3	G1S4
Semantic_Zoom_Levels	elegant	elegant	elegant	elegant
View_Information_Object	confusing	precise	precise	precise
Resize_Behvior	*empty*	competent	competent	competent
ViewModel	inconvenient	convenient	convenient	convenient
Model	easy	easy	easy	easy
Drag_Drop	pleasant	pleasant	pleasant	pleasant
InformationLandscape	beautiful	beautiful	beautiful	beautiful
SurfaceHandler	*empty*			
DBServer	complicated	complicated	complicated	confusing
RootCollection	good	good	good	good
LandscapeHandler		good	good	good
UserFunctions		beautiful	beautiful	beautiful
Commands		convenient	convenient	convenient
VisualProperties		*empty*	precise	precise
MouseInput		inconvenient	easy	easy
MouseHandler		inconvenient	inconvenient	inconvenient
SurfaceInput			easy	easy
DataBackend			*empty*	competent
RotateBehavior				

Step 3 – Visualizing changes over time:

While the above analysis is in principle also possible by looking at the original maps, this next part of the analysis requires the graphML based digital representations. This allows us to use graph analysis software to further break down and analyze the links between nodes. As we are especially interested in changes over time, we find animations to be particularly useful, as has also been demonstrated in the literature (Heer & Robertson, 2007). The graph analysis

research project *visone*,[11] which can be downloaded and used for free for non-commercial uses, provides the necessary functionality. It easily enables creation of an animation between two or several graphs and highlights any changes. For example, nodes are animated as they move to a new position, new nodes are smoothly faded in, disappearing links are marked in red before fading out, and new links are marked in green before becoming permanent. When analyzing one group in detail, this is very helpful. We recommend using the results from Step 2 as a focus point for the eye; then, play back and forth between the maps several times to identify the details. To obtain even more comprehensible animations when comparing two groups with each other or groups with the master map, another useful operation is available: namely, the automatic dynamic graph layout. This is helpful because each group (as well as the master map), while perhaps semantically similar, may have very different spatial layouts that can make visual comparisons difficult. *visone* employs a framework for offline dynamic graph drawing, meaning that all states of a graph are known before a layout is to be computed, as is the case here. The underlying layout algorithm used is the energy-based technique *stress minimization* (Gansner, Koren, & North, 2004), which generally produces better results than comparable energy-based techniques and also scales very well (Brandes U. P., 2008). In a dynamic graph layout, the objective is to preserve the mental map of a viewer – i.e., the parts of a layout where the graph does not change much should not be altered over the course of time, therefore producing coherent layouts and facilitating easy comparison between successive states. However, layout quality in terms of faithful representation of structural features in the graph and maintaining dynamic stability are naturally opposing objectives in most cases. The algorithm employed in *visone* explicitly models this trade-off with an *anchoring*-approach (Brandes & Wagner, 1997) (Frishman & Tal, 2008), penalizing point-wise deviations of a node's position from a reference position during layout calculation. A stability parameter $0<=\alpha<=1$ allows control between quality and stability. Using

[11] http://www.visone.info (online 30.07.2011)

α=0 corresponds to regular stress minimization for each individual layout, whereas α=1 will result in the reference layout for each state.

Regarding the reference layout, there are three options available. We can use either one of the input graphs as a reference, which is a sensible choice for comparisons with the master map or for comparisons of different groups at one point in time: take the previous state as a reference for the current one, or compute an aggregated layout of the whole sequence as a reference, which worked best for comparing a series of graphs from one group.

Figure 44 shows the original map for group 5 after the last session as well as the master map. The main problem for a visual analysis here is that the layout of the maps is completely different. Looking at the maps, we cannot really see whether the same concepts are being used and whether they are connected similarly or completely differently. The layout algorithm makes this a much easier task, as it rearranges the group map using the master map as a reference (see Figure 45). Afterwards, the two maps share the same layout and we can easily spot several differences but also similarities by looking at it. The lower part of the graph stays more or less completely stable (the prototype concepts are missing in the master map). The upper part seems similar as well, but the animation reveals some differences. The *commands* concept is missing and the connected *usefunctions of an object* concept is incorrectly connected directly to the *view* concept. This indicates that the *commands* were treated as a black box and usage could be enhanced and simplified with templates or code snippets. In addition, several *attached behavior* concepts are missing from group 5's map. Since the functionality existed in the prototype, they probably took advantage of this by copying existing code without understanding the underlying conceptual model. This could cause problems when new behaviors must be designed that are not provided by the framework.

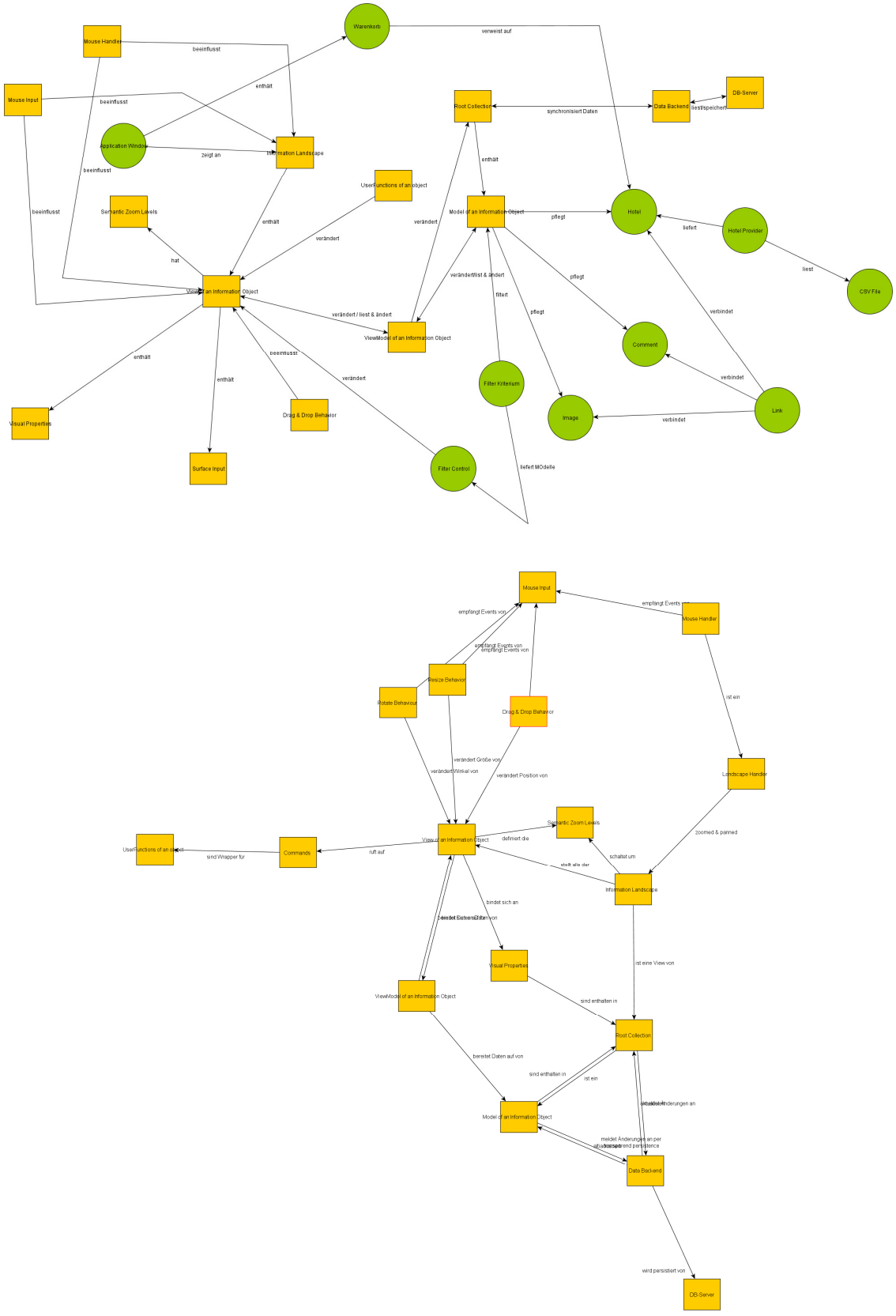


Figure 44: Top: original group 5 map, bottom: master map

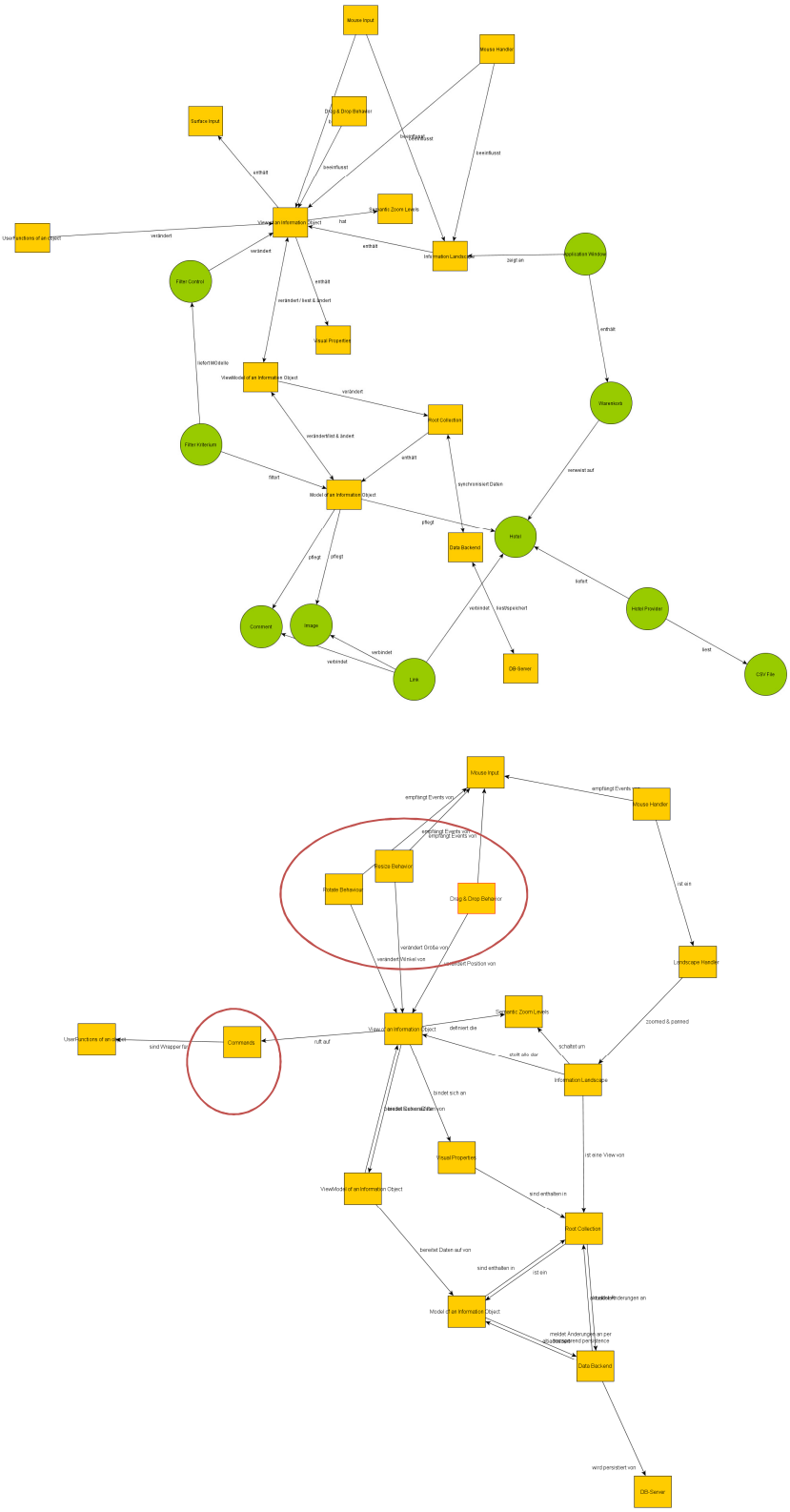


Figure 45: Top: group 5 map based on the stress minimization layout and the master map as reference, bottom: master map (α = 75%). Red markings: some of the differences between the maps.

Step 4/0 – Video analysis:

We intentionally have not discussed video analysis before this point, as this step is not really method-specific. Nevertheless, analyzing the videotaped sessions can reveal insights that are difficult to identify from the result-based analysis shown here. Participants often discuss the position and the linking of concepts in detail; sometimes, they argue about it, which obviously should be considered when analyzing the data in detail. If time constraints do not allow detailed video analysis, note-taking during the sessions can also help to identify important situations. In any case, the video data should be used to verify any claims.

By analyzing the video material in detail, one can also gain a better understanding of how people actually learn an API. While we have not analyzed the material from our case studies with respect to this question, such analysis could provide insight for better API design guidelines that would facilitate learning.

4.4.4 Case Study Conclusion

We were able to identify three main issues with the ZOIL API with the help of the Concept Maps method. First, people have difficulty understanding the concept of different input handlers. Video analysis revealed that they misused the concepts because they expected a different functionality based on their prior experiences. Second, the MVVM pattern, while causing less trouble than expected, still led to some misconceptions and was widely rated with negative adjectives. In some cases, concepts that should have connected to the View or the Model were connected to the ViewModel, indicating that users had problems clearly separating these concepts from one another. Third, we observed several group-specific issues with individual concepts that did not cause problems on a general level. The insight gained here will nevertheless help to create a more usable interface. The duration of four sessions also was shown to be appropriate, as the concept maps had mostly converged by the third session.

4.5 Discussion

In this section, we will discuss several dimensions of the method that have either shown to be controversial or that need clarification.

4.5.1 Usability vs. Learnability vs. Mental Models

We have described a variety of integrated data-gathering means and analysis possibilities, such as the map itself, the adjectives, the problem areas, and the development over time. However, one might wonder what it actually is that we are investigating here. Is it really about usability, or is it about learnability? Or is it a visualization for mental models? The answer to this question is not trivial, as the method can be adjusted to focus on any one of these three issues.

First, the map itself is certainly an approach to visualizing the mental model of the user. A “broken” mental model by itself is not necessarily a reason for concern, since we as users often have incorrect mental models of technology and still manage quite well (see, for example, Payne’s early studies on mental models of ATMs (Payne, 1991)). However, what the mental model allows us to do is to compare it among users and with the API developers. Therefore, it is important to consider the API developers “master” map not as the perfect map but simply as an additional perspective on the API. Differences do not necessarily mean that users have a usability problem, but it gives us reason to investigate further as to why these differences exist and what the consequences could be. With respect to learnability, this is definitely one of the main goals of the method in terms of identifying learning barriers and understanding how they develop over time, whether they disappear, and why. We use a longitudinal design specifically to address extended learning, as Grossmann would call it (Grossman, Fitzmaurice, & Attar, 2009). However, this does not mean that the method does not capture issues that are beyond learnability; rather, the method captures “immediate” problems as well, such as those identified through problem areas. The longitudinal design itself allows us to distinguish between learnability issues and those that persist over time and may be eventually considered usability is-

sues. Cross-sectional studies in particular are incapable of providing this insight. In addition, as Saldaña also pointed out (Saldaña, Analyzing longitudinal qualitative observational data, 2008), it is important to look not only for changes, but also for things that remain stable. The Concept Maps approach allows easy visual identification of both of these aspects.

4.5.2 Elicitation vs. Construction

An issue related to the idea of visualizing the mental model of the users is the question of whether the approach actually elicits the mental model or whether it is an act of construction that may or may not have similarities with the mental model of the user. As the Concept Maps approach is a qualitative and indirect method, we must first state that these two almost always go together. When we examine artifacts from users, whether they are concept maps or texts from interviews, they are always to some extent constructed during the data-gathering session and are not pure reflections of the users' minds. An interview question can trigger a completely new thinking process; asking users to create a concept map of an API certainly requires them to think about the API in a way they most certainly have not done before. Even with standardized approaches such as the Experience Sampling Method, the authors state that they cannot guarantee whether the method actually captures immediate emotions or whether the emotions are partly constructed because of the measurement instrumentation (Hektner, Schmidt, & Czikszentmihalyi, 2007).

What is important for the Concept Maps approach is that construction is actually a valuable part of the method. The constructive process is predominant in the first session, and it allows us to get much more insight into the understanding of the users, as the process of creating the map is itself iterative and no simple reproduction of an existing mental image. This means that participants do not simply create a map from the start to the end of a session; instead, they discuss it, think about it, try out different layouts and connections, and include many iterative steps overall. This construction phase is itself a valuable resource for the researcher to understand the thought process of the users. It also may help

participants to become aware of certain issues they might otherwise forget to mention.

However, elicitation becomes an important part of the longitudinal design. During subsequent sessions, participants start looking at the API differently, and because they know what is expected from them in the concept-mapping sessions, they can prepare themselves mentally. We often observed that in later sessions, participants were actively scanning their map for aspects that no longer matched their perceptions. At that point, less construction is involved than elicitation.

Especially in order to be able to analyze change processes, we think it is important that the construction proportion is reduced in favor of elicitation. Construction always introduces a potential variability and therefore an error term. When comparing maps over time, it is important that changes in the map can be ascribed to a changed understanding of the users. This also explains why we did not propose allowing users to create a completely new map in each session – this would result in a constructive process every time, making it much more difficult to ascribe changes between maps to changes in the understanding of the users, as the constructive process itself could have introduced a bias. The Concept Maps approach essentially tries to benefit as much as possible from the constructive processes that are inherent to qualitative and indirect research methods while also benefitting from elicitation in the longitudinal design.

4.5.3 Comparison to Other Methods

In our case studies, we have not explicitly compared our approach to existing techniques; i.e., we did not run a control group with usability tests to see how the results would differ. The reason is that we do not regard the Concept Maps approach as a competitor for these approaches, but rather as a technique that can provide a completely different and complementary perspective. The longitudinal design itself captures a dimension that is not possible with cross-sectional designs. In addition, the concept maps provide a more abstracted

view that allows insight into obvious usability issues but also potential usability issues that might not result in broken code during the study. We have discussed some examples in the case study presented, such as the input handler problem, which we would regard as a potential usability issue. Capturing these potential issues is difficult with usability tests or code inspection, as these are more focused on the resulting code artifact; in contrast, Concept Maps try to focus more on an "understanding" artifact.

The method also should scale very well for use together with other methods, such as code inspection or usability testing (e.g., pre/post of the longitudinal design). While the concept map is the integral part discussed at length here, the process itself can also reveal many issues (the discussions between participants, requests for help from the API expert, etc.); using the method in a cross-sectional design could also be helpful as well (although many benefits would thereby be lost).

4.5.4 Costs of the Method

We consider the method as very flexible in regard to costs. There are several aspects that can be regarded as cost factors but also others that allow the reduction of costs. One important cost factor is the skill level of the participants. This is a problem inherent to API usability, as not everyone is suitable to be a participant in such a study. Using an API requires certain skills and programming expertise; APIs are somewhat expert systems and no one would expect that a person without any programming experience could make use of them (although there are some tools that focus on end-user programming, an area in which ease of use gets more important again). As the Concept Maps method is an approach in which participants are required to work on a real or realistic task in their work environment or at home for a realistic time period, the method is potentially more expensive than usability tests in which participants only have to focus on the study during the test session. Therefore, it may be important to find a win-win situation for such tests. This can be reached in several ways. In our case, we were able to integrate the study within a lecture, so that participants

received grades for the prototype they were developing with the API. Such an approach is obviously only possible at educational institutions. Depending on the maturity of the API and the environment in which it is being developed, there could be several other possibilities. For example, one could organize programming contests with prizes, so that participants would 1) work on the same task and 2) be motivated to complete the task. In some cases, it might be possible to combine the work with an API with productive work that must be done regardless. This may be more suitable post-release of an API, but for certain user groups, a beta release may be sufficient for productive applications.

Another cost factor is the researcher who administers the concept-mapping sessions and guides the analysis of the data. In a best-case scenario, this person would have API experience and could understand the specific problems. However, an API expert is needed as well to assist in designing the study and to be involved during data analysis. In the best case, he or she would also attend the concept-mapping sessions to be able to provide feedback when needed.

From a cost-reduction perspective, important aspects are the hands-on materials that are used and the easy-to-recreate tool chain. Others are the study's scope, duration, and integration with other research methods, which can all be easily adjusted. The method can be defined to capture very detailed parts of the API or a general picture. The study itself can be trimmed to a few days with only 2-3 data-gathering sessions or even be used in a cross-sectional design, accompanying a usability test, for example. Obviously, this would eliminate the identification of learning barriers and the elicitation part of the process, as discussed above. We feel that having such an abstracted representation in addition to more outcome-related techniques, such as code inspection or usability testing, could lend much-needed balance to the analysis and interpretation of the data.

4.6 Conclusion

In this chapter, we have presented the Concept Maps method as a longitudinal approach to evaluating the usability of an API. The method is based on the idea that concept maps can be used to elicit and assess the knowledge users have of complex and abstract domains – for example, science (as in the original use of concept maps) or an API (as in our case). We have shown that the high-level view above code-level that the Concept Maps demand and provide can make it easier to recognize misconceptions and usability issues before they lead to serious problems after deployment. The method provides a variety of means of data gathering and analysis, such as the possibility to rate concepts or indicate problem areas. The graph-based structure of the maps allows creation of digital representations of the maps, which facilitates the use of graph analysis tools such as *visone.* Using the Concept Maps method in a cross-sectional design can greatly increase the benefit of an API usability test, for example. By allowing participants to create a personal map and extend and modify it over time, changes in understanding become visible to the researcher and learning barriers can be observed.

While we purely used graph analysis tools such as *visone* to rearrange the layout, animate transitions and thereby allow the “manual” discovery of similarities and differences, these tools could also allow the application of similarity algorithms. Thereby, means of measuring the level of agreement between participants and the API developers could be achieved. The difficulty here lies in the definition of the similarity measure, as multiple dimensions play a role and these are interdependent. Imagine concept A should be connected directly to concept B and concept C – however one person connects A only with B and B with C. Another person connects A with C and C with B. In both cases, there is one correct and one wrong connection. However, these two variants might be very different substantially and this has to be taken into account for such a similarity measure.

The method can also be applied in a more realistic task-setting than what is possible in a usability test. While we have focused on the creation of the maps, they can also serve as helpful prompts during interviews, allowing participants to spatially locate problems; this was greatly appreciated in our studies. Last but not least, the Concept Map method can help participants gain a better understanding of the API, as it asks them to reflect on their usage; concept maps have proven to be useful learning aids in the past. While this certainly influences the method itself (as is true for many evaluation approaches), we see this as being of specific benefit as a training opportunity for participants who come from within an organization that is developing an API for internal use and who are meant to be end-users as well. Finally, asking the API developers to create a master map can also help in identifying potential issues upfront.

In the future, it will be interesting to investigate how the method could also be combined with theoretical frameworks, such as Clarke's approach of using the cognitive dimensions. It might also be interesting to investigate in detail the effect of using pre-defined vs. user-defined concepts. While we have comprehensively discussed how to use the Concept Maps method and the various possibilities regarding analysis of the data, we think that one significant benefit of the method is its flexibility in terms of materials and data-gathering techniques. Finally, it opens up an enormous design space for future research on how to elicit knowledge and understanding of an API, which could be beneficial for both analyzing the usability of one specific API and for designing future APIs.

5 Summary & Conclusion

Empirical research is one of the central elements of Human-Computer Interaction as it is essential to two main research areas, 1) to validate novel concepts and interaction designs and 2) to gain more insight into human's behavior and interaction with technology. Therefore, methodological advances regarding the research methods are needed that allow us to study today's phenomena. As HCI is a multi-disciplinary field, many of these research methods advances have been adaptions from other areas, such as the social sciences, psychology, marketing research, educational research, among others. However, only few have addressed the fundamental flaw of cross-sectional research – the inability to investigate change processes and capture the dynamics of human-computer interaction. Longitudinal Research is able to address this issue, as it gathers data over time and allows time-dependent analysis. In HCI, there are mainly two areas that already rely on longitudinal research in the broadest sense: 1) the area of input device design and evaluation, where learning has often been an obstacle to introducing new techniques to the broader public (think about the QWERTY keyboard design and its more effective alternatives); 2) the area of User Experience research, which has long admitted that human emotions are not stable and therefore longitudinal research is needed to capture a more complete picture of the User Experience (e.g. see the research by Karaponas, Hassenzahl and others cited in this thesis). However, as we have discussed in this thesis, Longitudinal Research is still seldom and for many people difficult to apply. What is still lacking to this day are a) a common understanding and framework for Longitudinal Research in HCI, that provides a basis for application and for discussion of methodological advances, b) more research investigating specific longitudinal methods and tool that support Longitudinal Research.

In this thesis we have addressed these issues by providing several contributions to the field of Longitudinal Research in HCI. In Chapter 2 we have developed and discussed a taxonomy for Longitudinal Research. We explored in detail the kind of research questions that can be addressed with longitudinal re-

search and the different research designs that are suitable to answer these questions. Thereby, we provided a common ground for Longitudinal Research in HCI which has been missing in the literature. The taxonomy enables the application of Longitudinal Research and should also facilitate the discussion for further methodological advances in this area. In terms of research questions, we have distinguished between three main research questions: 1) Interest in the average over time, 2) interest in the effect of change, and 3) interest in the process of change. We have shown that the first type of research questions is in principal identical to cross-sectional research and does not aim at analyzing or explaining changes. It still allows a better validation of the data, as the impact of time-dependent outliers is reduced. The second area is the simplest form of time-dependent research questions, as not the change process itself is studied but the effect and outcome, allowing the researcher to contrast data over time. This kind of longitudinal research is often applied when novel interaction designs are evaluated and e.g. learning is either a confounding variable (so we are interested in the point in time when there is no more learning) or we would like to show the effect of learning. The third area of research questions now is interested in the change process itself, in its form, potential lows and peaks, and the underlying reasons. These kinds of research questions are more difficult to address and more advanced analysis methods, both statistical and qualitative are needed here. Therefore, we also presented and discussed two specific analysis methods in detail, multi-level growth curve modeling and survival analysis. In retrospective, for some of the studies done in relation to this thesis, these methods could have provided additional insights. For example in the first laser-pointer experiment described in (Gerken, Bieg, Dierdorf, & Reiterer, 2009a), we were not able to model the exact amount of “spare” time between each data-gathering session, but simply used a “day” metric. As not every participant was able to come to the lab at the same time each day, this day metric was not stable across participants and over time. Multi-level growth curve modeling would have allowed us to integrate this into the model and help us explain more of the variance in the data. It would have also allowed us to test different relationships between the change measurement and time and not just assume a power law of learning. In principal, as the method does not assume the independence of

each measurement (as ANOVA requires), it is much better suited for statistical analysis in longitudinal studies in general.

In the remainder of this chapter, we have brought together these different research questions with research designs, analysis techniques, data-gathering schedules, and data-gathering techniques, providing the interested researcher with a comprehensive framework for conducting longitudinal research.

Based on the taxonomy, we were furthermore able to identify additional areas for research, namely in terms of longitudinal methods that are tailor-made to the challenge of capturing change processes in qualitative data over time and the tool support to allow researchers to capture longitudinal data in the field in a very flexible way. With regards to tool support, Chapter 3 discusses the diary method which we then implemented in PocketBee to provide researchers with a state of the art tool for multi-modal data-capturing in longitudinal field studies. PocketBee is based on the Android mobile platform and allows researchers to conduct ESM as well as diary studies up to very complex study designs and data-gathering schedules. In two studies we could show the usability of the interface that is provided to participants. Besides we have presented a conceptual design for a novel researcher interface, which allows the flexible configuration of study designs without any programming knowledge.

Chapter 4 addresses the issue of capturing change processes in qualitative longitudinal data over time by presenting the Concept Maps method in the specific use case of API usability. The approach allows researchers to let users externalize and visualize their mental representation of an API. The longitudinal design then makes changes over time explicit in the artifacts created, the concept maps, as participants are asked to update and correct this visual representation over time as they continue to explore and learn the API. We have presented a case study that shows the potential benefits of the approach which allows us to identify learning barriers in APIs as well as potential usability issues – aspects which do work in the implemented code but are misrepresented in the concept map. These can be a result of trial & error and the copying of example code, which is a common approach when working with an API. Such issues are

much more difficult to obtain with code-inspection or usability lab tests, as the possibility to cross-validate the data is missing. Nevertheless, it is important to point out the Concept Maps, as every other evaluation approach, should be complemented with other techniques. The approach itself, however, already offers a variety of possibilities to cross-validate data, as the maps include ratings, problem areas, the structure and connections which can be combined with code inspection and also analysis of the audio protocol during concept map sessions. Thereby, the approach provides a rich and very flexible set of data-gathering tools with the inherent advantage of allowing the analysis of qualitative longitudinal data.

Figure 46 provides an overview of the contributions of this thesis and how they relate to each other. Overall, we envision that this thesis helps in increasing the awareness for longitudinal research in HCI.

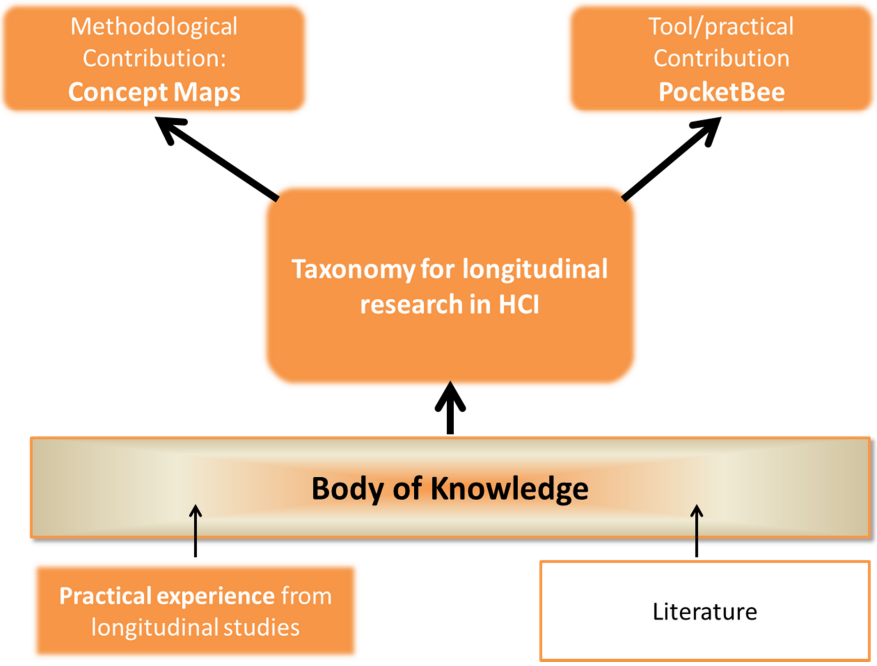


Figure 46: The individual contributions of this thesis (filled with orange)

As our goal was to provide a generic overview (Chapter 2: Taxonomy) accompanied with specific contributions (Chapter 3 & 4, PocketBee & Concept Maps), we are sure that there are a lot more research opportunities left in this area. Specifically, the inherent drawbacks of longitudinal research as discussed in the introduction need to be studied more in detail and if and to what extent they

specifically apply to HCI. Besides, and this is especially important, applying longitudinal research means to get accustomed to novel ways of data analysis, beyond ANOVA and grounded theory. In particular, we think that longitudinal research has to take the path to regard the data-collection even more in relation to the data-analysis. As the Concept Maps approach has shown, only this tight coupling allows for example an efficient analysis of change processes in qualitative data.

We are convinced that our research encourages others to undertake Longitudinal Research and enter the discussion about methods and approaches. The work presented in this thesis should foster such discussion and help novices to get accustomed to the field.

6 Postscript

„The journey is its own reward“

The path to writing this thesis has not been a straightforward process at all times. Similar to the topic of this thesis, changes have been a recurring theme. While the overall topic of the thesis eventually stayed the same, not every research that the author undertook along the way, found its way inside the numerous pages. Reasons for this are manifold, but eventually the author simply had too much fun chasing interesting projects, even in case that the relation to the thesis topic was not apparent. This chapter is meant to give a rough overview of some of these projects. All of them were done in collaboration with fellow PhD students or student researchers and might eventually (or already have) find their way into other theses. Therefore, this chapter is not meant to take any credit away from the tremendous work these people did but instead allow the reader to understand, which thesis-related or completely unrelated topics attracted the author of this thesis along the way. If applicable, the following sections simply reproduce the abstract or short summaries of the related publications.

6.1 Dynamic Text Filtering for Improving the Usability of Alphasliders on Small Screens (Büring, Gerken, & Reiterer, 2007)

Authors: Thorsten Büring, Jens Gerken, Harald Reiterer

Publication: IV '07: Proceedings of the Eleventh International Conference on Information Visualisation, IEEE Computer Society, Jul 2007

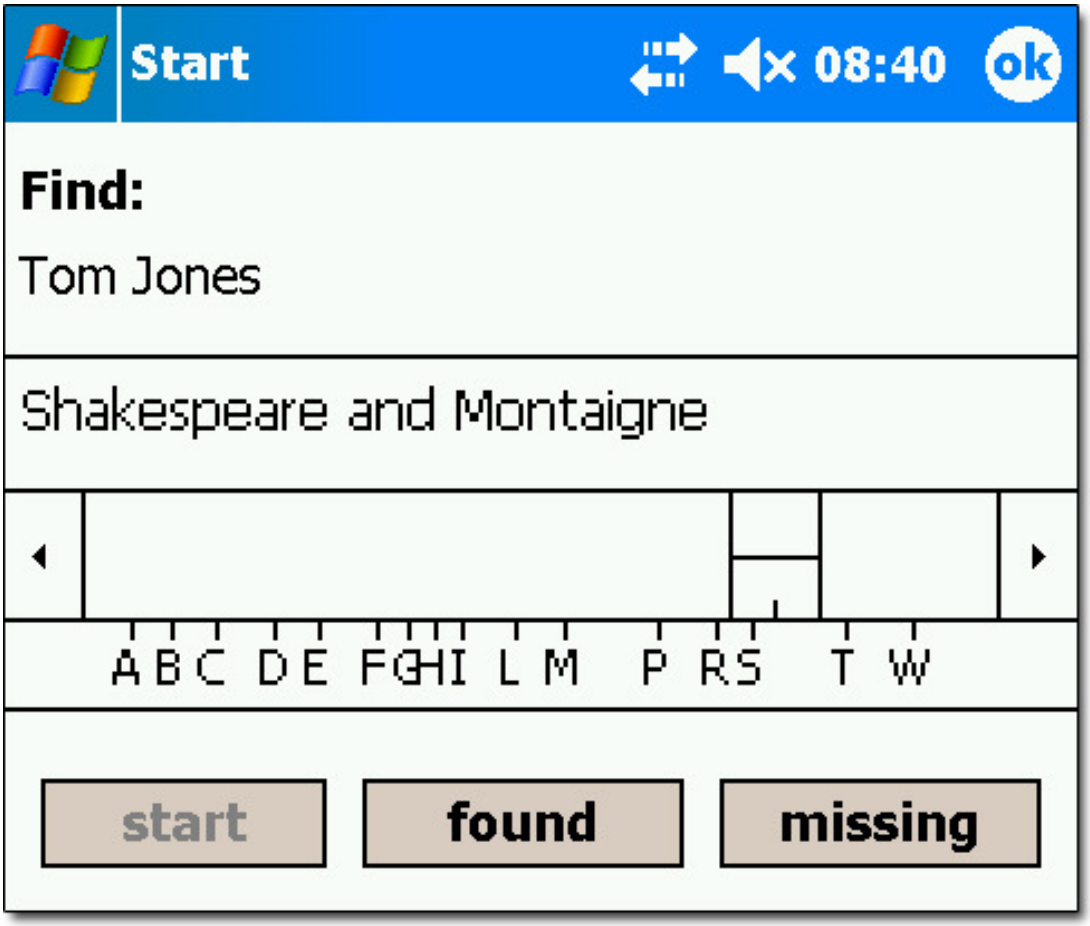


Figure 47: Alphaslider for mobile devices

Previous research has shown that Alphasliders are an effective tool for searching an alphabetically sorted list when only limited screen space is available for the graphical user interface. To improve user satisfaction, we propose equipping the widget with a novel text filter to dynamically limit the slider range (see Figure 47). In this way, users are supported in locating target items and in iden-

tifying records that are missing. The results of a comparative user evaluation run on a Personal Digital Assistant showed that 8 out of 12 participants preferred the filter widget to the classic interface. We further suggest an enhanced Alphaslider design to speed up user interaction.

6.2 Blockbuster – A Visual Explorer for Motion Picture Data (Rexhausen, et al., 2007)

Authors: Sebastian Rexhausen, Mischa Demarmels, Hans-Christian Jetter, Mathias Heilig, Jens Gerken, Harald Reiterer

Publication: INFOVIS 07: IEEE Visualization 2007 Conference Compendium, IEEE Computer Society, (Awarded with 3rd place), Nov 2007.

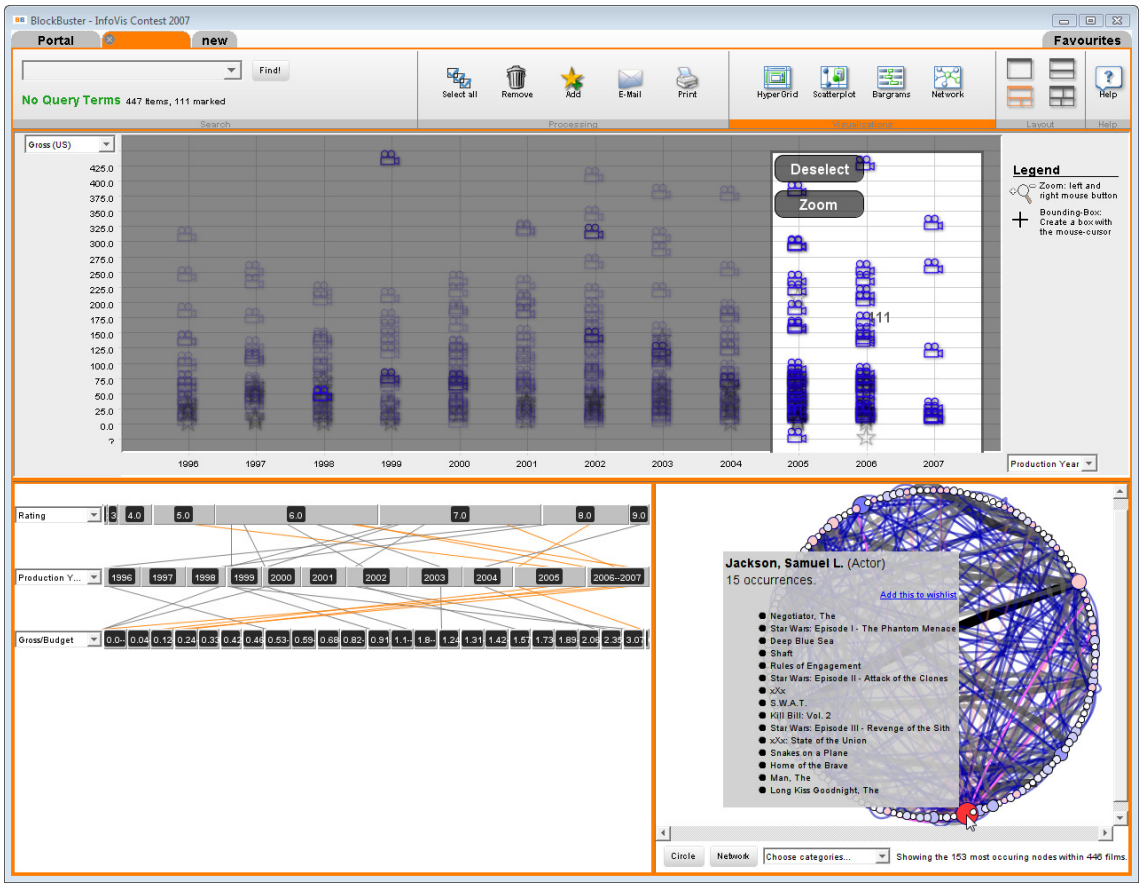


Figure 48: Blockbuster - A Visual Explorer for Motion Picture Data

In this project we introduced our visual explorer “Blockbuster” as a contribution to the InfoVis Contest 2007 (see Figure 48). The system’s development followed a user-centered design process and a design rationale considering not only the pragmatic qualities of the system, but also hedonic qualities like aesthetics or “joy-of-use”. Apart from briefly outlining the employed visualization techniques, we will focus on Blockbuster’s interaction design, which is aimed at

facilitating the selection, combination and mutual filtering of visualizations under a consistent interaction paradigm. Blockbuster thereby demonstrates the potential of information visualization for end-user-centered applications that blur the boundaries of information visualization, visual information seeking and browsing.

6.3 Zoom interaction design for pen-operated portable devices (Büring, Gerken, & Reiterer, 2008)

Authors: Thorsten Büring, Jens Gerken, Harald Reiterer

Publication: International Journal of Human-Computer Studies, Elsevier, vol. 66 (2008), p. 605-627.

Figure 49: The test setup for the zoom-interaction experiment

Maps are currently the most common application domain for zoomable user interfaces (ZUIs). Standard techniques for controlling such interfaces on pen-operated devices usually rely on sequential interaction, i.e. the users can either zoom or pan. A more advanced technique is speed-dependent automatic zooming (SDAZ), which combines rate-based panning and zooming into a single operation and thus enables concurrent interaction. Yet another navigation strategy is to allow for concurrent, but separate, zooming and panning. However, due to the limitations of stylus input, this feature requires the pen-operated device to be enhanced with additional input dimensions. We propose one unimanual approach based on pen pressure, and one bimanual approach in which users pan

the view with the pen while manipulating the scale by tilting the device. In total, we developed four interfaces (standard, SDAZ, pressure, and tilting) and compared them in a usability study with 32 participants (see Figure 49). The results show that SDAZ performed well for both simple speed tasks and more complex navigation scenarios, but that the coupled interaction led to much user frustration. In a preference vote, the participants strongly rejected the interface and stated that they found it difficult and irksome to control. This result enhances previous research, which in most cases found a high user preference for SDAZ, but focused solely on simple speed tasks. In contrast, the pressure and tilt interfaces were much appreciated, which, considering the novelty of these approaches, is highly encouraging. However, in solving the test tasks the participants took hardly any advantage of parallel interaction. For a map view of 600 x 600 pixels, this resulted in task completion times comparable to those for the standard interface. For a smaller 300 x 300 pixels view, the standard interface was actually significantly faster than the two novel techniques. This ratio is also reflected in the preference votes. While for the larger 600 x 600 pixels view the tilt interface was the most popular, the standard interface was rated highest for the 300 x 300 pixels view. Hence, on a smaller display, precise interaction may have an increased impact on the interface usability. Overall, we believe that the alternative interaction techniques show great potential for further development. In particular, a redesign should encourage parallel interaction more strongly and also provide improved support for precise navigation.

6.4 Adaptive Pointing – Design and Evaluation of a Precision Enhancing Technique for Absolute Pointing Devices (König, Gerken, Dierdorf, & Reiterer, 2009)

Authors: Werner A. König, Jens Gerken, Stefan Dierdorf, Harald Reiterer

Publication: Interact 2009: Proceedings of the the twelfth IFIP conference on Human-Computer Interaction, Springer, Uppsala, Sweden, p. 658-671, Aug 2009

Figure 50: The Adaptive Pointing technique in combination with a laser-pointer as input device in front of a Powerwall

We present Adaptive Pointing, a novel approach to addressing the common problem of accuracy when using absolute pointing devices for distant interaction (see Figure 50). First, we discuss extensively some related work concerning the problem-domain of pointing accuracy when using absolute or relative pointing devices. As a result, we introduce a novel classification scheme to more clearly discriminate between different approaches. Second, the Adaptive Pointing technique is presented and described in detail. The intention behind this approach is to improve pointing performance for absolute input devices by implicitly adapting the Control-Display gain to the current user‟s needs without violating users‟ mental model of absolute-device operation. Third, we present an experiment comparing Adaptive Pointing with pure absolute pointing using a laser-

pointer as an example of an absolute device. The results show that Adaptive Pointing results in a significant improvement compared with absolute pointing in terms of movement time (19%), error rate (63%), and user satisfaction.

6.5 Lessons Learned from the Design and Evaluation of Visual Information Seeking Systems (Gerken, et al., 2009b)

Authors: Jens Gerken, Mathias Heilig, Hans-Christian Jetter, Sebastian Rexhausen, Mischa Demarmels, Werner A. König, Harald Reiterer

Publication: International Journal on Digital Libraries, vol. 10, no. 2, p. 49--66, Springer Verlag, Aug 2009.

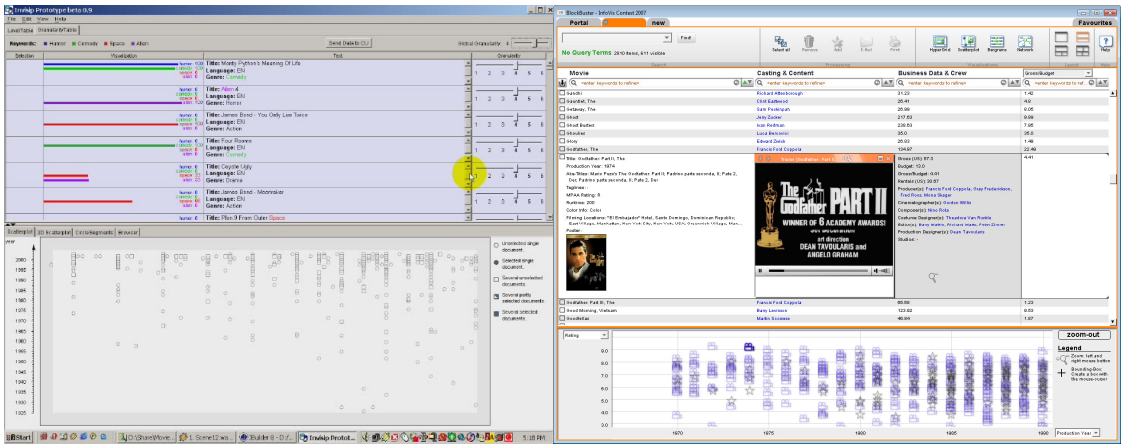


Figure 51: Visual information seeking systems from different times - VisMeb (left) and Mediovis (right)

Designing information-seeking systems has become an increasingly complex task as today's information spaces are rapidly growing in quantity, heterogeneity, and dimensionality. The challenge is to provide user interfaces that have a satisfying usability and user experience even for novice users. Although information visualization and interaction design offer solutions, many information seeking systems such as online catalogs for libraries or web search engines continue to use outdated user-interface concepts developed decades ago. In this paper, we will present four principles that we identified as crucial for the successful design of a modern visual information-seeking system. These are (1) to support various ways of formulating an informationneed, (2) to integrate analytical and browsing-oriented ways of exploration (seeFigure 51), (3) to provide views on different dimensions of the information space, and (4) to make search a pleasurable experience. These design principles are based on our experience

over a long period in the user-centered design and evaluation of visual information-seeking systems. Accordingly, we will showcase individual designs from our own work of the past 10 years to illustrate each principle and hence narrow the gap between the scientific discussion and the designing practitioner that has often hindered research ideas from becoming reality. However, most of the times search is only one part of a higher level user activity (e.g. writing a paper). Thus future research should focus on the challenges when regarding search in such a broader context. We will use the final two chapters to point out some of these challenges and outline our vision of an integrated and consistent digital work environment named Zoomable Object-oriented Information Landscape (ZOIL).

6.6 Can "touch" get annoying? (Gerken J. , Jetter, Schmidt, & Reiterer, 2010c)

Authors: Jens Gerken, Hans-Christian Jetter, Toni Schmidt, Harald Reiterer

Publication: In Proceedings of ITS 2010: The ACM International Conference on Interactive Tabletops and Surfaces 2010, Poster Session.

Figure 52: A participant explaining a mechanical instrument on paper

While touch interaction with tabletops is now widely accepted as a very natural and intuitive form of input, only little research has been carried out to understand whether and how it might interfere with our natural ways of gestural communication. This poster presents a study that aims at understanding the importance of touching physical and virtual artifacts during discussion or collaboration around a table (see Figure 52). Furthermore, it focuses on how users compensate for conflicts between non-interactivity and interactivity created by unintended touch interaction when using a multi-touch enabled tabletop. In our study, we asked participants to explain illustrations of technical or physical mechanisms, such as the workings of an airplane wing. We observed whether and how they used gestures to do so on a touch sensitive Microsoft Surface tabletop and on a sheet of paper. Our results suggest that touching is an essential part of such an activity and that the compensation strategies people adapt to avoid conflicts may reduce precision of communication and increase the physical strain on the user.

6.7 Materializing the Query with Facet-Streams – A Hybrid Surface for Collaborative Search on Tabletops (Jetter H.-C. , Gerken, Zöllner, Reiterer, & Milic-Frayling, 2011)[12]

Authors: Hans-Christian Jetter, Jens Gerken, Michael Zöllner, Harald Reiterer, Natasa Milic-Frayling

Publiction: CHI'11: Proceedings of the 29th international conference on Human factors in computing systems, ACM Press, May 2011, honorable mention award).

[12] The author of this thesis presented this topic at an invited talk at Microsoft Research Cambridge, UK: http://research.microsoft.com/apps/video/default.aspx?id=147152 (online April 2011)

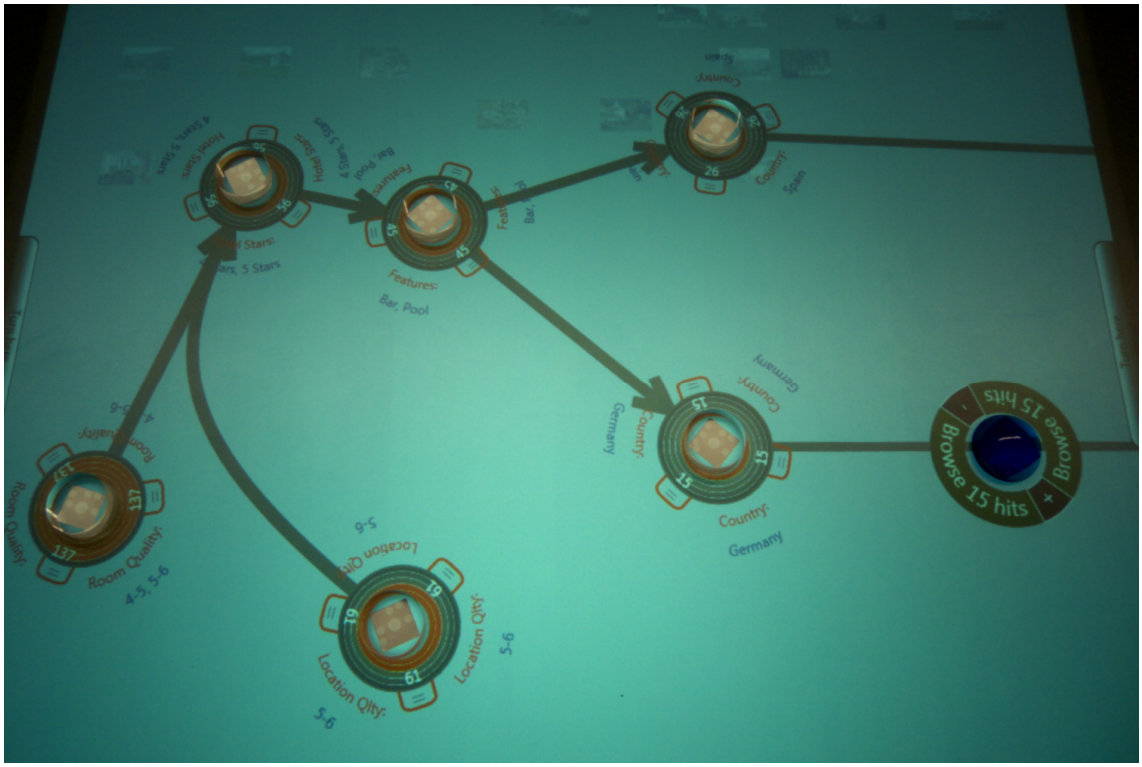


Figure 53: Facet-Streams hybrid interface. Top: Facet-Token Interaction, Bottom: complex query with Boolean logic

We introduce “Facet-Streams”, a hybrid interactive surface for co-located collaborative product search on a tabletop (see Figure 53). Facet-Streams combines techniques of information visualization with tangible and multi-touch interaction to materialize collaborative search on a tabletop. It harnesses the expressive power of facets and Boolean logic without exposing users to complex formal notations. Two user studies reveal how Facet-Streams unifies visual and tangible expressivity with simplicity in interaction, supports different strategies and collaboration styles, and turns product search into a fun and social experience[13].

[13] A video illustrating Facet-Streams in detail is available here: http://www.youtube.com/watch?v=giDF9lKhCLc (online April 2011)

6.8 Hidden Details of Negotiation: The Mechanics of Reality-Based Collaboration in Information Seeking (Heilig, et al., 2011)

Authors: Mathias Heilig, Stephan Huber, Jens Gerken, Mischa Demarmels, Katrin Allmendinger, Harald Reiterer

Publication: to appear in INTERACT 2011: Proceedings of 13th IFIP TC13 Conference on Human-Computer Interaction

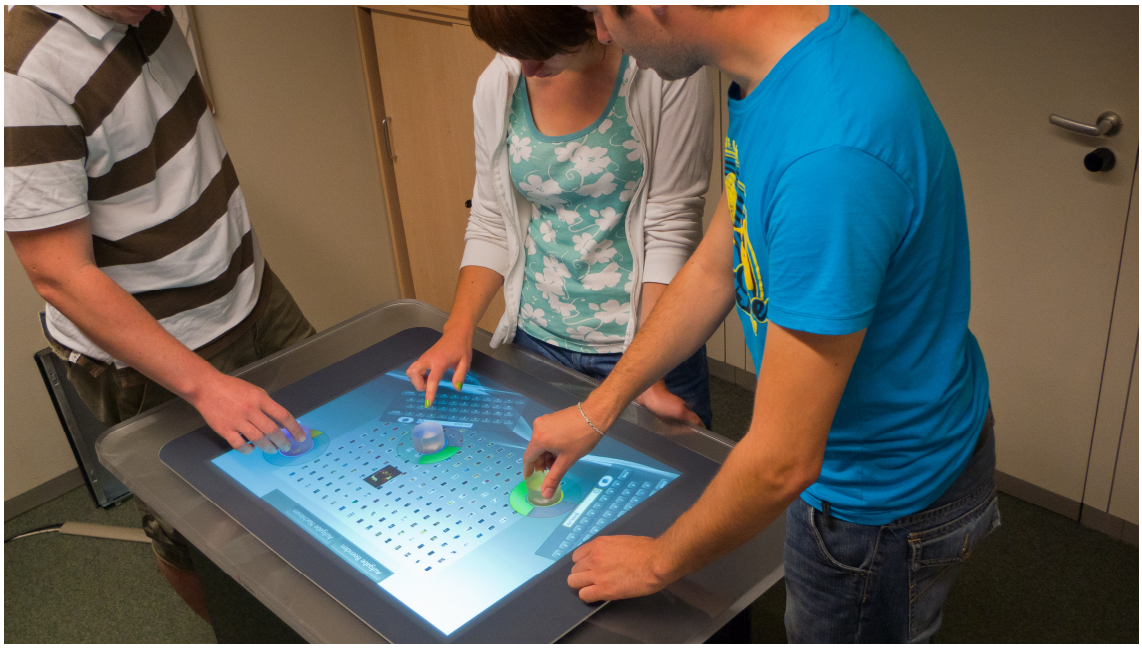


Figure 54: Three persons interacting with the tangible search interface

Social activities such as collaborative work and group negotiation can be an essential part of information seeking processes. However, they are not sufficiently supported by today's information systems as they focus on individual users working with PCs. Reality-based UIs with their increased emphasis on social, tangible, and surface computing have the potential to tackle this problem. By blending characteristics of real-world interaction and social qualities with the advantages of virtual computer systems, they inherently change the possibilities for collaboration, but until now this phenomenon has not been explored sufficiently. Therefore, this paper presents an experimental user study

that aims at clarifying the impact such reality-based UIs and its characteristics have on collaborative information seeking processes. Two different UIs have been developed for the purpose of this study. One is based on an interactive multi-touch tabletop in combination with on-screen tangibles (see Figure 54), therefore qualifying as a reality-based UI, while the other interface uses three synchronized PCs each controlled by keyboard and mouse. A comparative user study with 75 participants in groups of three was carried out to observe fundamental information seeking tasks for co-located collaboration. The study shows essential differences of emerging group behavior, especially in terms of role perception and seeking strategies depending on the two different UIs.

References

Alaszewski, A. (2006). *Using diaries for social research.* Sage Publications Ltd.

Anick, P., & Kantamneni, R. G. (2008). A Longitudinal Study of Real-time Search Assistance Adoption. *SIGIR 2008.* ACM Press.

Ballagas, R., Memon, F., Reiners, R., & Borchers, J. (2007). iStuff mobile: rapidly prototyping new mobilephone interfaces for ubiquitous computing. *Proc. CHI '07* (pp. 1107-1116). ACM Press.

Barksdale, J., & McCrickard, D. (2010). Concept Mapping in Agile Usablity: A Case Study. *Proc. CHI 2010 EA*, (pp. 4691-4694).

Barrett, L. F., & Barrett, D. J. (2001). An Introduction to Computerized Experience Sampling in Psychology. *Social Science Computer Review*(19), pp. 175-185.

Beaton, J., Myers, B., Stylos, J., Jeong, S., & Xie, Y. (2008). Usabiltiy evaluation for enterprise SOA APIs. *Proc. of SDSOA '08* (pp. 29-34). ACM Press.

Bergman, O., Beyth-Marom, R., Nachmias, R., Gradovitch, N., & Whittaker, S. (2008). Improved Search Engines and Navigation Preference in Personal Informatoin Management. *ACM Transactions on Information Systems, 26*(4), pp. 20:1-20:24.

Bieg, H.-J. (2008). *Laserpointer and Eye Gaze Interaction - Design and Evaluation.* Master-Thesis, University of Konstanz, Computer-Science, Konstanz.

Bijleveld, C. C., & van der Kamp, L. J. (1998). *Longitudinal Data Analysis: Designs, Models, and Methods.* London: Sage Publications Ltd.

Bleich, C. (1999). *Veränderungen der Paarbeziehungsqualität vor und während der Schwangerschaft sowie nach der Geburt des ersten Kindes.* Stuttgart: Reichle, B.; Werneck, H.

Bloch, J. (2005). How to write a good API and why it matters. *LCSD workshop at OOPSLA.*

Bogen II, P. L., Francisco-Revialla, L., Furuta, R., Hubbard, T., Karadkar, U. P., & Shipman, F. (2007). Longitudinal Study of Changes in Blogs. *JCDL 2007.* ACM Press.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annual review of psychology, 54*(1), pp. 579-616.

Borchers, J. (2001). *A Pattern Approach to Interaction Design.* Wiley.

Borgman, C. (2003). Designing digital libraries for usability. In B. Buttenfield, & V. Houase (Eds.), *Digital library use: social practice in design and evaluation.* Cambridge: ACM.

Brandes, U. P. (2008). An experimental study on distance-based graph drawing. *Proc. 16th Int. Symp. on Graph Drawing* (pp. 218-229). Springer.

Brandes, U., & Wagner, D. (1997). A Bayesian paradigm for dynamic graph layout. *Proc. 5th Int. Symp. on Graph Drawing* (pp. 236-247). Springer.

Broderick, J. E., & Stone, A. A. (2006). Paper and Electronic Diaries: Too Early for Conclusions on Compliance Rates and Their Effects - Comment on Green, Rafaeli, Bolger, Shrout, and Reis (2006). *Psychological Methods, 11*(1), pp. 106-111.

Bruun, A., Gull, P., Hofmeister, L., & Stage, J. (2009). Let Your Users do the Testing: AComparison of Three Remote Asynchronous Usability Testing Methods. *CHI 2009* (pp. 1619-128). ACM Press.

Büring, T., Gerken, J., & Reiterer, H. (2007). Dynamic text filtering for improving the usability of alphasliders on small screens. *Information Visualization 2007 (IV).* IEEE.

Büring, T., Gerken, J., & Reiterer, H. (2008). Zoom interaction design for pen-operated portable devices. *International Journal of Human-Computer Studies*(66), pp. 605-627.

Cairns, P., & Cox, A. L. (Eds.). (2008). *Research Methods for Human-Computer Interaction.* Cambridge University Press.

Cantor, D. (2008). A review and summary of studies on panel conditioning. In S. Menard, *Handbook of Longitudinal Research: Design, Measurement, and Analysis* (pp. 123-138). Elsevier.

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*, pp. 601-613.

Carter, S., & Mankoff, J. (2005). When Participants Do the Capturing: The Role of Media in Diary Studies. *CHI 2005* (pp. 899-908). ACM Press.

Carter, S., Mankoff, J., & Heer, J. (2007). Momento: support for situated ubicomp experimentation. *in Proc. of CHI'07.* ACM Press.

Castellucci, S. J., & MacKenzie, I. S. (2008). Graffiti vs. Unistrokes: An Empirical Comparison. *CHI 2008.* ACM Press.

Clarke, S. (2004, May). Measuring API Usability. *Dr. Dobbs Journal*, 6-9.

Courage, C., Jain, J., & Rosenbaum, S. (2009). Best Practices in Longitudinal Research: A CHI 2009 Workshop. *CHI 2009: Extended Abstracts.* ACM Press.

Courage, C., Rosenbaum, S., & Jain, J. (2008). Exploring Best Practices in Longitudinal Usability Studies: A UPA 2008 Workshop. *UPA 2008.*

Cutter, C. (1876). *Rules for a Dictionary Catalog.* Washington, D.C.: Government Printing Office.

Cwalina, K., & Abrams, B. (n.d.). *Framework design guidelines.* (Addison-Wesley, Ed.) 2005.

Czerwinski, M., Horvitz, E., & Wilhite, S. (2004). A Diary Study of Task Switching and Interruptions. *In Proc. of CHI'04.* ACM Press.

Daughtry, J., Farooq, U., Myers, B., & Stylos, J. (2009, July). API Usability: Report on Special Interest Group at CHI. *Software Engineering Notes*.

Daughtry, J., Stylos, J., Farooq, U., & Myers, B. (2009). API Usability: CHI'2009 Special Interest Group Meeting. *Proc. CHI'09 EA.* ACM Press.

Day, J. F. (2006). Evaluating Web Lectures: A Case Study from HCI. *CHI 2006: Extended Abstracts.* ACM Press.

Dayton, C. M. (2008). An introduction to latent class analysis. In S. Menard, *Handbook of Longitudinal Research: Design, Measurement, and Analysis* (pp. 357-372). Elsevier.

de Souza, C., Redmiles, D., Cheng, L., Millen, D., & Patterson, J. (2004). Sometimes You Need to See Through Walls - A Field Study of Application Programming Interfaces. *Proc. CSCW*, (pp. 63-71).

Dierdorf, S. (2011). *Spezifikation von Anforderung und Interaktionsdesign einer multi-modalen Tagebuch-Anwendung für Feld-Studien.* Master-Thesis, University of Konstanz, Computer Science.

Douglas, S., Kirkpatrick, A., & MacKenzie, I. S. (1999). Testing pointing device performance and user assessment with the ISO 9241, Part 9 standard. *CHI 1999* (pp. 215-222). ACM Press.

Ducheneaut, N., Yee, N., Nickell, E., & Moore, R. J. (2006). "Alone Together?" Exploring the Social Dynamics of Massively Multiplayer Online Games. *CHI 2006.* ACM Press.

Ellis, B., Stylos, J., & Myers, B. (2007). The Factory Pattern in API Design: A Usability Evaluation. *Proc. ICSE '07* (pp. 302-312). ACM Press.

Eppler, M. (2006, June). A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. *Information Visualization*(5), pp. 202-210.

Farooq, U., & Zirkler, D. (2010). API Peer Reviews: A Method for Evaluating Usability of Application Programming Interfaces. *Proc. CHI 2010*, (pp. 207-210).

Feng, J., Zhu, S., Hu, R., & Sears, A. (2008). Speech Technology in Real world Environment: Early Results from a Long Term Study. *ASSETS 2008* (pp. 233-234). ACM Press.

Friess, E. (2008). Defending Design Decisions with Usability Evidence: A Case Study. *CHI 2008 Extended Abstracts* (pp. 2009-2016). ACM Press.

Frishman, Y., & Tal, A. (2008). Online dynamic graph drawing. *IEEE Trans. on Visualiz. and Comp. Graphics, 14*(4), pp. 727-740.

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. (2007). MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. *Mobisys: Proceedings of the 5th international conference on Mobile systems, applications and services.* ACM Press.

Gansner, E., Koren, Y., & North, S. (2004). Graph drawing by stress majorization. *Proc. 12th Int. Symp. on Graph Drawing* (pp. 239-250). Springer.

Garzonis, S., Jones, S., Jay, T., & O'Neill, E. (2009). Auditory Icon and Earcon Mobile Service Notifications: Intuitiveness, Learnability, Memorability and Preference. *CHI 2009* (pp. 1513-1522). ACM Press.

Gerken, J., Bak, P., & Reiterer, H. (2007). Longitudinal Evaluation Methods in Human-Computer Studies and Visual Analytics. *Metrics for the Evaluation of Visual Analytics (InfoVis 2007 Workshop).*

Gerken, J., Bak, P., Jetter, H.-C., Klinkhammer, D., & Reiterer, H. (2008a). How to use interaction logs effectively for usability evaluation. *BELIV 2008: Beyond time and errors (A CHI 2008 Workshop).* ACM Press.

Gerken, J., Demarmels, M., Dierdorf, S., & Reiterer, H. (2008b). HyperScatter – Modellierungs- und Zoomtechniken für Punktdiagramme. *Mensch & Computer 2008: Viel mehr Interaktion, 8. Konferenz für interaktive und kooperative Medien.* Oldenbourg Verlag.

Gerken, J., Bieg, H.-J., Dierdorf, S., & Reiterer, H. (2009a). Enhancing Input Device Evaluation: Longitudinal Approaches. *CHI 2009: Extended Abstracts.* ACM Press.

Gerken, J., Heilig, M., Jetter, H.-C., Rexhausen, S., Demarmels, M., König, W. A., & Reiterer, H. (2009b, August). Lessons Learned from the Design and Evaluation of Visual Information Seeking Systems. *International Journal on Digital Libraries, 10*(2), pp. 49-66.

Gerken, J., & Reiterer, H. (2009c). Eine Taxonomie für Längsschnittstudien in der MCI. *Mensch & Computer 2009.* Oldenbourg Verlag.

Gerken, J., Jetter, H.-C., & Reiterer, H. (2010a). Using Concept Maps to Evaluate the Usability of APIs. *CHI 2010: Extended Abstracts.* ACM Press.

Gerken, J., Dierdorf, S., Schmid, P., Sautner, A., & Reiterer, H. (2010b). PocketBee: a multi-modal diary for field research. *nordiCHI: In Proc. of the 6th Nordic Conference on Human-Computer Interaction.* ACM Press.

Gerken, J., Jetter, H.-C., Schmidt, T., & Reiterer, H. (2010c). Can "touch" get annoying? *In Proceedings of ITS 2010: The ACM International Conference on Interactive Tabletops and Surfaces 2010, Poster Session.* ACM Press.

Gerken, J., Jetter, H.-C., Zöllner, M., Mader, M., & Reiterer, H. (2011). The Concept Maps Method as a Tool to Evaluate the Usability of APIs. *CHI'11: Proceedings of the 29th international conference on Human factors in computing systems.* ACM Press.

González, V., & Kobsa, A. (2003). A Workplace Study of the Adoption of Information Visualization Systems. *I-KNOW'03: 3rd International Conference on Knowledge Management*, (pp. 92-102).

Gorlenko, L. (2005). Long-Term (Longitudinal) Research and User Experience Design. *UPA Conference 2005 - Advanced Topic Seminar.*

Green, T., & Petre, M. (1996). Usability Analysis of Visual Programming Environments: A Cognitive Dimensions Framework. *Journal of Visual Languages and Computing, 7*(2), pp. 131-174.

Grossman, T., Fitzmaurice, G., & Attar, R. (2009). A survey of software learnability: metrics, methodologies and guidelines. *CHI 2009.* ACM Press.

Hammer, S., Leichtenstern, K., & André, E. (2010). Using the mobile application EDDY for gatherin user information in the requirement analysis. *EICS: In*

Proc. of the 2nd ACM SIGCHI Symposium on Engineering Interactive Computing Systems. ACM Press.

Harada, S., Wobbrock, J. O., Malkin, J., Bilmes, J. A., & Landay, J. A. (2009). Longitudinal Study of People Learning to Use Continuous Voice-Based Cursor Control. *CHI 2009* (pp. 347-356). ACM Press.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. (P. Hancock, & N. Meshkati, Eds.) *Human Mental Workload*.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. *Mensch & Computer 2003.* B.G. Teubner.

Heer, J., Card, S., & Landay, J. (2005). prefuse: a toolkit for interactive information visualization. *Proc. CHI '05.* ACM Press.

Heer, J., & Robertson, G. (2007). Animated Transitions in Statistical Data Graphics. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 13*(6), pp. 1240-1247.

Heilig, M., Huber, S., Gerken, J., Demarmels, M., Allmendinger, K., & Reiterer, H. (2011). Hidden Details of Negotiation: The Mechanics of Reality-Based Collaboration in Information Seeking. *Interact 2011.* ACM Press.

Hektner, J. M., Schmidt, J. A., & Czikszentmihalyi, M. (2007). *Experience Sampling Method: Measuring the qulity of everyday life.* Sage Publications Ltd.

Hilbe, J. M., & Hardin, J. W. (2008). Generalized estimating equations for longitudinal panel analysis. In S. Menard, *Handbook of Longitudinal Research: Design, Measurement, and Analysis* (pp. 467-474). Elsevier.

Hinze, A., Chang, C., & Nichols, D. (2010). Contextual queries express mobile information needs. *MobileHCI 2010.* ACM Press.

Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., . . . Wood, K. (2006). SenseCam: a Retrospective Memory Aid. *Ubicomp'06: In Proc. of the 8th International Conference on Ubiquitous Computing.* ACM Press.

Hofte, G. H. (2007). What's that Hot Thing in my Pocket? SocioXensor, a smartphone data collector. *Proc. of e-Social Science: Third International Conference on e-Social Science.*

Hoggan, E., & Brewster, S. A. (2010). CrossTrainer: Testing the Use of Multimodal Interfaces in Situ. *CHI 2010* (pp. 333-342). ACM Press.

Holleis, P., Wagner, M., Böhm, S., & Koolwaaij, J. (2010). Studying Mobile context-aware Social Services in the Wild. *NordiCHI 2010* (pp. 207-216). ACM Press.

Howard, S., Kjeldskov, J., Skov, M. B., Garnoes, K., & Grünberger, O. (2006). Negotiating Presence-in-Abscence: Contact, Content and Context. *CHI 2006* (pp. 909-912). ACM Press.

Jain, J., Ghosh, R., & Dekhil, M. (2008). Multimodal Capture of Consumer Intent in Retail. *CHI 2008: Extended Abstracts* (pp. 3207-3212). ACM Press.

Jain, J. (2010). InfoPal: A System for Conductin and Analyzing Multimodal Diary Studies. *UPA 2010: Proceedings of the Usability Professional's Association International Conference.* UPA.

Jain, J., Rosenbaum, S., & Courage, C. (2010). SIG: Best Practices in Longitudinal Research. *CHI 2010: Extended Abstracts.* ACM Press.

Jensen, K. L. (2009). RECON: Capturing mobile and ubiquitous interaction in real contexts. *mobileHCI 2009.*

Jetter, H.-C., Gerken, J., König, W. A., Grün, C., & Reiterer, H. (2005). HyperGrid - Accessing Complex Information Spaces. *HCI UK 2005: People and Computers XIX - The Bigger Picture, Proceedings of the 19th British HCI Group Annual Conference 2005.* Springer.

Jetter, H.-C., Gerken, J., Zöllner, M., & Reiterer, H. (2010). Model-based Design and Prototyping of Interactive Spaces for Information Interaction. *Proc. of Human-Centred Software Engineering (HCSE).*

Jetter, H.-C., Gerken, J., Zöllner, M., Reiterer, H., & Milic-Frayling, N. (2011). Materializing the Query with Facet-Streams – A Hybrid Surface for Collaborative Search on Tabletops. *CHI'11: Proceedings of the 29th international conference on Human factors in computing systems.* ACM Press.

Kahnemann, D., Krueger, A., Schkade, D., Schwarz, N., & Stone, A. A. (2004, December). A survey method of characterizing daily life experiences: The Day Reconstruction Method. *Science*, pp. 1776-1780.

Kaptelinin, V., & Czerwinski, M. (2007). *Beyond the desktop metaphor: Designing integrated digital work environments.* Cambridge: MIT Press.

Karapanos, E., Martens, J.-B., & Hassenzahl, M. (2009). On the Retrospective Assessment of Users' Experiences Over Time: Memory or Actuality? *CHI 2009: Extended Abstracts.* ACM Press.

Karapanos, E., Martens, J.-B., & Hassenzahl, M. (2009). Reconstructing Experiences through Sketching. *CoRR, abs/0912.5343*.

Karapanos, E., Zimmermann, J., Forlizzi, J., & Martens, J.-B. (2009). User Experience Over Time: An Initial Framework. *CHI 2009* (pp. 729-733). ACM Press.

Karapanos, E., Zimmermann, J., Forlizzi, J., & Martens, J.-B. (2010, September). Measuring the dynamics of remembered experience over time. *Interacting with Computers, 22*(5), pp. 328-335.

Karlson, A. K., Iqbal, S. T., Meyers, B., Ramos, G., Lee, K., & Tang, J. C. (2010). Mobile taskflow in context: a screenshot study of smartphone usage. *CHI 2010.* ACM Press.

Khan, V. J., Markopoulos, P., Eggen, B., Ijsselsteijn, W., & Ruyter, de, B. (2008). ReconExp: A way to reduce the data loss of the experience sampling method. *MobileHCI'08: In Proc. of the 10th International Conference on Human-Comuter Interaction with mobile devices and services.* ACM Press.

Khan, V., Markopoulos, P., & Eggen, B. (2009). Features for the future Experience Sampling Tool. *MobileHCI'09.*

Kirstensson, P. O., & Denby, L. C. (2009). Text Entry Performance of State of the Art Unconstrained Handwriting Recognition: A Longitudinal User Study. *CHI 2009* (pp. 567-570). ACM Press.

Kjeldskov, J., Skov, M. B., & Stage, J. (2005). Does Time Heal? A Longitudinal Study of Usability. *OZCHI 2005.* ACM Press.

Klemmer, S., Lie, J., Lin, J., & Landay, J. (2004). Papier-Maché: toolkit support for tangible input. *Proc. CHI ' 04.* ACM Press.

Ko, A., Myers, B., & Aung, H. (2004). Six learning barriers in end-user programming systems. *Proc. IEEE Symp. on Visual Languages and Human-Centric Computing* (pp. 199-206). IEEE.

König, W. A., Gerken, J., Dierdorf, S., & Reiterer, H. (2009). Adaptive Pointing – Design and Evaluation of a Precision Enhancing Technique for Absolute Pointing Devices. *Interact 2009: Proceedings of the the twelfth IFIP conference on Human-Computer Interaction.* Springer.

König, W. A., Rädle, R., & Reiterer, H. (2010, Feb). Interactive Design of Multimodal User Interfaces - Reducing technical and visual complexity. *Journal on Multimodal User Interfaces, 3*(3), pp. 197-213.

Krishnamurthy, B., & Wills, C. E. (2009). Privacy Diffusion on the Web: A Longitudinal Perspective. *WWW 2009* (pp. 541-550). ACM Press.

Lazar, J., Feng, D. J., & Hochheiser, D. H. (2009). *Research Methods in Human-Computer Interaction.* John Wiley and Sons.

Lichtner, V., Kounkou, A. P., Dotan, A., Kooken, J. P., & Maiden, N. A. (2009). An online forum as a user diary for remote workplace evaluation of a work-integrated learning system. *CHI 2009: Extended Abstracts* (pp. 2955-2970). ACM Press.

Lira, E., Ripoll, P., Peiro, J., & V., O. (2008). How do different types of intragroup conflict affect group potency in virtual compared with face-to-face teams? A longitudinal study. *Behaviour & Information Technology, 27*(2), pp. 107-114.

Liu, N., Ying, L., & Wang, X. (2010). Data Logging plus E-diary: towards an Online Evaluation Approach of Mobile Service Field Trial. *mobileHCI 2010* (pp. 287-290). ACM Press.

Luke, D. A. (2008). Multilevel growth curve analysis for quantitative outcomes. In S. Menard, *Handbook of Longitudinal Research: Design, Measurement, and Analysis* (pp. 545-564). Elsevier.

MacKenzie, I. S., & Zhang, S. X. (1999). The Design and Evaluation of a High-Performance Soft Keyboard. *CHI 1999* (pp. 25-31). ACM Press.

MacKenzie, I. S., Kauppinen, T., & Silfverberg, M. (2001). Accuracy measures for evaluating comuter pointing devices. *CHI 2001* (pp. 9-16). ACM Press.

Majaranta, P., Ahola, U.-K., & Spakov, O. (2009). Fast Gaze Typing with an Adjustable Dwell Time. *CHI 2009* (pp. 357-360). ACM Press.

McClure, J., Sonak, B., & Suen, H. (1999). Concept Map Assessment of Classroom Learning: Reliability, Validity, and Logistical Practicality. *Journal of Research in Science Teaching, 36*(4), pp. 475-492.

McLachlan, P., Munzner, T., Koutsofios, E., & North, S. (2008). LiveRAC: Interactive Visual Exploration of System Management Time-Series Data. *CHI 2008* (pp. 1483-1491). ACM Press.

McLellan, S., Roesler, A., Tempest, J., & Spinuzzi, C. (1998). Building more usable APIs. *IEEE Software, 15*(3), pp. 78-86.

Menard, S. (2002). *Longitudinal Research* (2. Edition ed.). London: Sage Publications Ltd.

Menard, S. (Ed.). (2008). *Handbook of Longitudinal Research: Design, Measurement, and Analysis.* Elsevier.

Menard, S. (2008). Panel analysis with logistic regression. In S. Menard, *Handbook of Longitudinal Research: Design, Measurement, and Analysis* (pp. 505-522). Elsevier.

Mendoza, V., & Novick, D. G. (2005). Usability over Time. *SIGDOC* (pp. 151-158). ACM Press.

Myers, B., Hudson, S., & Pausch, R. (2000). Past, present, and future of user interface software tools. *ACM Trans. Computer-Human Interaction, 7*(1), pp. 3-28.

Novak, J., & Gwon, D. (1984). *Learning How to Learn.* (C. U. Press, Ed.) Cambridge, UK.

Nylander, S., Lundquist, T., Brännström, A., & Karlson, B. (2009). "It's just easier with the phone" - a diary study of Internet access from cell phones. *Pervasive 2009.*

Palmer, C., Zavalina, O., & Mustafoff, M. (2007). Trneds in Metadata Practices: a Longitudinal Study of Collection Federation. *JCDL 2007.* ACM Press.

Patterson, G. R. (2008). Orderly change in a stable world: The antisocial trait as a chimera. In S. Menard (Ed.), *Handbook of Longitudinal Research: Design, Measurement, and Analysis* (pp. 153-166). Elsevier.

Payne, S. J. (1991). A descriptive study of mental models. *Behavior & Informatoin Technology, 10*(3), pp. 3-21.

Perlin, K., & Fox, D. (1993). Pad: an alternative approach to the computer interface. *SIGGRAPH 1993* (pp. 57-64). ACM Press.

Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., & Aucinas, A. (2010). EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. *Ubicomp'10: Proc. of the12th ACM International Conference on Ubiquitous Computing.* ACM Press.

Reis, H. T., & Wheeler, L. (1991). Studying Social Interaction with the Rochester Interaction Record. *Advances in Experimental Social Psychology, 24*, pp. 269-311.

Rexhausen, S., Demarmels, M., Jetter, H.-C., Heilig, M., Gerken, J., & Reiterer, H. (2007). Blockbuster - A Visual Explorer for Motion Picture Data. *InfoVis 2007: Contest Entry.* IEEE.

Rieger, M. (2009). *Novel Input Devices:Technophobia, Practice, and Acceptance.* Konstanz, Germany: University of Konstanz.

Rieman, J. (1993). The diary study: A workplace-oriented research tool to guide laboratory efforts. *Proc. INTERCHI '03.* ACM Press.

Rogers, Y., Sharp, H., & Preece, J. (2007). *Interaction Design: Beyond Human-Computer Interaction.* Wiley.

Saldaña, J. (2003). *Longitudinal Qualitative Research: Analyzing Change Through Time.* Walnut Creek, California: AltaMira Press.

Saldaña, J. (2008). Analyzing longitudinal qualitative observational data. In S. Menard, *Handbook of Longitudinal Research: Design, Measurement, and Analysis* (pp. 297-312). Elsevier.

Saraiya, P., North, C., Lam, V., & Duca, K. (2006). An Insight-based Longitudinal Study of Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics, 12*.

Schmidt, P. (2011). *Improving Data-Gathering in Field Studies by Using Electronic Devices.* Master-Thesis, University of Konstanz, Computer Science.

Seay, A. F., & Kraut, R. E. (2007). Project Massive: Self-Regulation and Problematic Use of Online Gaming. *CHI 2007.* ACM Press.

Shneiderman, B., & Plaisant, C. (2006). Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. *Proc. BELIV '06, a AVI 2006 workshop.* ACM Press.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis.* New York: Oxford University Press.

Sporka, A. J., Kurniawan, S., Mahmud, M., & Slavik, P. (2007). Longitudinal Study of Continuous Non-Speech Operated Mouse Pointer. *CHI 2007: Exended Abstracts* (pp. 2669-2674). ACM Press.

Strauss, A. (1987). *Qualitative analysis for social scientists.* Cambridge: Cambridge University Press.

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes OVer Time: The Psychology of Panel Conditioning. In P. Lynn (Ed.). Wiley.

Sung, J., Christensen, H. I., & Grinter, R. E. (2009). Robots in the Wild: Understanding Long-term Use. *HRI 2009* (pp. 45-52). ACM Press.

Taris, T. W. (2000). *A primer in longitudinal data analysis.* London: SAGE Publications.

Teevan, J., Dumais, S. T., Liebling, D. J., & Hughes, R. L. (2009). Changing How People View Changes on the Web. *UIST 2009* (pp. 237-246). ACM Press.

Trumble, K., & Marshall, R. (2003). *The Library of Alexandria.* New York: Clarion Books.

Tuisku, O., Majaranta, P. I., & Räihä, K.-J. (2008). Now Dasher! Dsah Away! Longitudinal Study of Fast Text Entry by Eye Gaze. *ETRA 2008* (pp. 19-26). ACM Press.

Tulach, J. (2008). *Practical API Design: Confessions of a Java Framework Architect.* (Apress, Ed.)

Urbina, M. H., & Huckauf, A. (2010). Alternatives to Single Character entry and Dwell Time Selection on Eye Typing. *ETRA 2010* (pp. 315-322). ACM Press.

van der Kleij, R., Paashuis, R., & Schraagen, J. M. (2005, January 25). On the passage of time: temporal differences in video-mediated and face-to-face interaction. *International Journal of Human-Computer Studies, 62*, pp. 521-542.

Vaughan, M. W., & Dillon, A. (2006, December 27). Why structure and genre matter for users of digital information: A longitudinal experiment with readers of a web-based newspaper. *International Journal of Human-Computer Studies, 64*, pp. 502-526.

Vaughan, M., & Courage, C. (2007). SIG: Capturing Longitudinal Usability: What really affeects user performance over time. *CHI 2007: Extended Abstracts.* ACM Press.

Vaughan, M., Beale, R., Courage, C., Hammontree, M., Jain, J., Rosenbaum, S., & Welsh, D. (2008). Longitudinal Usability Data Collection: Art versus Science? *CHI 2008: Extended Abstracts.* ACM Press.

Voida, S., & Mynatt, E. D. (2009). "It Feels Better Than Filing": Everyday Work Experiences in an Activity-Based Computing System. *CHI 2009* (pp. 259-268). ACM Press.

von Wilamowitz-Moellendorf, M., Hassenzahl, M., & Platz, A. (2007). Veränderung in der Wahrnehmung und Bewertung interaktiver Produkte. *Mensch & Computer 2007* (pp. 49-58). Oldenbourg Verlag.

Waterton, J., & Lievesley, D. (1989). Evidence of conditioning effects in the British Social Attitudes Panel Survey. *Panel Surveys*, pp. 319-339.

Wheeler, L., & Reis, H. (1991). Self-Recording of Everyday Life Events: Origins, Types, and Uses. *Journal of Personality, 59*(3), pp. 339-354.

White, R. W., & Drucker, S. M. (2007). Investigating Behavioral Variability in Web Search. *WWW 2007* (pp. 21-30). ACM Press.

White, R. W., & Horvitz, E. (2009). Cyberchondira: Studies of the Escalation of Medical Concerns in Web Search. *ACM Transactions on Information Systems, 27*(4), pp. 23:1-23:37.

Wigdor, D., Penn, G., Ryall, K., Esenther, A., & Shen, C. (2007). Living with a Tabletop: Analysis and Observations of Long Term Office Use of a Multi-Touch Table. *IEEE International Workshop on Horizontal Interactive Human-Computer Systems* (pp. 60-67). IEEE.

Wilson, M. L., & mc schraefel. (2008). A Longitudinal Study of Exploratory Keyword Search. *JCDL 2008* (pp. 52-55). ACM Press.

Wobbrock, J. O., Myers, B. A., & Rothrock, B. (2006). Few-key Text Entry Revisited: Mnemonic Gestures on Four Keys. *CHI 2006* (pp. 489-492). ACM Press.

Yeh, R. B., Liao, C., Klemmer, S. R., Guimbretière, F., Lee, B., Kakaradov, B., . . . Paepcke, A. (2006). ButterflyNet: A mobile Capture and Access System for Field Biology Research. *CHI 2006.* ACM Press.

Appendix A

This appendix gives an overview of those reviewed papers that fulfilled the requirements of being published at a major HCI conference or journal and that provided the necessary details to replicate the study design (42 in total). They serve is one of the foundations of the taxonomy presented in Chapter 2. The table shows which research questions were addressed and some additional details that should help researchers to quickly find a potentially related study. Thereby, this Appendix should serve as some reference guide for longitudinal studies.

The table does not include the research designs taxonomy, simply because nearly all studies implement a Panel design. Those who did not, feature this in the “Notes” column. More advanced designs, such as Retrospective Panels, are simply not common in HCI, yet.

Most studies address several types of research questions. We marked with green those types we identified as predominant in the study and with a lighter green those which were also addressed.

Title of the paper	Reference	Type of Study and data-gathering methods (e.g. experiment, ethnographic study, interviews, etc.)	**RQ: Average**	**RQ: Outcome**	**RQ: Pre-Post Comparisons**	**RQ: Shape**	**RQ: Event Occurence**	**RQ: In-depth**	Lab or field	Duration	Notes
A longitudinal study of exploratory and keyword search	(Wilson & mc schraefel, 2008)	Interaction logs, online forum, weekwise retrospective questionnaire, telephone interviews							Field	4 weeks	Descriptive statistics of averages across participants. Descriptive statistics of change from first visit to second website visit. Qualitative feedback also on average level and with respect to differences in first and second visit.
A Longitudinal Study of Real-time Search Assistance Adoption	(Anick & Kantamneni, 2008)	Log study							Field	4 months (4 waves)	Log files are analyzed over four sessions and the differences between sessions is analyzed with descriptive statistics (no graphical visualization)
A workplace study of the adoption of information visualization systems	(González & Kobsa, 2003)	Interviews							Field	Ca. 6 weeks	The study investigates how experts use information visualization in their real work. It analyzed how this bevavior changed over time and why. Therefore, it focused on qualitative interview data, but there is no information about the systematic extraction of changes in the data.
Alone Together? Exploring the Social Dynamics of Massively Multiplayer Online Games	(Ducheneaut, Yee, Nickell, & Moore, 2006)	Log Study							Field	Ca. 1 year	The study analyzes WoW. Interestingly, it uses not “time” as time-variable but the character level. It then analyzes playing time in relation to this character level both graphically and statistically.
Alternatives to Single Character Entry and Dwell Time Selection on Eye Typing	(Urbina & Huckauf, 2010)	Experiment with wpm and error measures							Lab	Unclear	RM-ANOVA with session as factor. Comparing of groups with mean over last three sessions (interest in outcome). Many descriptive statistics and graphs. Shows the possibilities of RM ANOVA for longitudinal data analysis. Unfortunately, no discussion of problems. Lots of graphical analysis to show shape of change.
An Insight-Based Longitudinal Study of Visual Analytics	(Saraiya, North, Lam, & Duca, 2006)	Observations + research diary							Field	Not stated	An in-depth study of how experts use information visualization and how this developed over time. The focus is on when and how insights were derived through the visualizations and what were the reasons.
Auditory Icon and Earcon Mobile Service Notifications: Intuitiveness, Learnability, Memorability and Preference	(Garzonis, Jones, Jay, & O'Neill, 2009)	Two lab experiments, a field study, web-based experiment							Lab + Field	Ca. 1 week	Very good study design that combines field and lab. The field part was hardly analyzed however. Individual growth modeling could have helped here. Basically, most analysis were done to compare pre-post or pre-middle-post. RM Anova comparing pre-post (lab 1 vs lab 2). Event occurence tests were done until all participants reached the event. Therefore, simple t-test were sufficient to compare these numbers.

Title of the paper	Reference	Type of Study and data-gathering methods (e.g. experiment, ethnographic study, interviews, etc.)	**RQ: Average**	**RQ: Outcome**	**RQ: Pre-Post Comparisons**	**RQ: Shape**	**RQ: Event Occurence**	**RQ: In-depth**	Lab or field	Duration	Notes
Changing how people view changes on the web.	(Teevan, Dumais, Liebling, & Hughes, 2009)	Feedback reports (positive, negative, neutral) (diary) + retrospective interviews							Field	2 weeks	Although the paper claims to analyze how the tool changed people's behavior, there is no real systematic analysis of change and not within the course of the study. It is reduced to retrospective reports or assessments of participants how it changed their behavior. Some descriptive statistics but mainly qualitative episodes on how the tool was used or how it "changed" the way people used the web.
CrossTrainer: Testing the Use of Multimodal Interfaces in Situ	(Hoggan & Brewster, 2010)	Lab based session, free usage in the field with tasks, 2 days for each variant							Field + lab	8 days	Nice study, analysis limited to descriptive and simple pre-post stuff. Would be interesting to see whether learning was different among conditions. Still, quite similar slopes from the graphics. Descriptive statistics for shape and event occurrence (+graphical). 2-factor ANOVA to compare first with last session for each modality. No statistical analysis of shape or comparison of shapes between modalities. Preferences were analyzed without time reference – simply averages.
Cyberchondria: Studies on the Escalation of Medical Concerns in Web Search	(White & Horvitz, Cyberchondira: Studies of the Escalation of Medical Concerns in Web Search, 2009)	Log-based field study, independent survey with retrospective questions							Field	11 months	Shape of change: identification of three, pre-determined classifications. Then mostly average analysis. Quantitative: Descriptive, ANOVA for effect of change pattern, graphical for distance of events (similar to survival methods).
Data Logging plus e-Diary: towards an Online Evaluation Approach of Mobible Service Field Trial	(Liu, Ying, & Wang, 2010)	E-diary + logging							Field	2 months	Basically, this paper presents a method and a case study. However, no real change analysis, although they claim to do so in p3, they then aggregate again over certain time periods (one day, one week). Descriptive statistics (frequencies) to analyze behavior (time dependent, but aggregated again – comparing weekdays with each other) Overall averages and descriptive statistics.
Defending Design Decisions with Usability Evidence: A Case study	(Friess, 2008)	Field observations and coding of events							Field	12 months	Longitudinal study with interest in the average person – however good example as so far, as data was checked for changes over time at least fundamentally (correlation between number of session and event occurance). Discourse Analysis: coding of incidents in different categories. Then, purely quantitative, descriptive analysis. Mainly averages over all participants are given and their range across sessions. It is stated that no correlation with session number is present.
Does time heal? A Longitudinal study of Usability	(Kjeldskov, Skov, & Stage, 2005)	Usability Test							Lab	18 months (2 data-gathering waves)	See section 2.2.3.2 Pre-Post Comparisons for a detailed discussion

Title of the paper	Reference	Type of Study and data-gathering methods (e.g. experiment, ethnographic study, interviews, etc.)	**RQ: Average**	**RQ: Outcome**	**RQ: Pre-Post Comparisons**	**RQ: Shape**	**RQ: Event Occurence**	**RQ: In-depth**	Lab or field	Duration	Notes
Evaluating Web Lectures: A Case Study from HCI	(Day, 2006)	Quasi-experiment							Field (classroom)	15 weeks	The study compared two user groups, one using web lectures and one using traditional lectures. Groups were compared overall and at distinct points during the semester (the exams + questionnaires). Simple ANOVA comparisons were conducted. Problematic: No Baseline measurement so experiment suffers the typical problems from repeated cross-sectional designs.
Fast Gaze Typing with an Adjustable Dwell Time	(Majaranta, Ahola, & Spakov, 2009)	Lab-based experiment							Lab	10 days	Only graphical analysis of change shape. ANOVA pre-post for size of effect. No analysis of curve itself. Again, the shape itself is merely described and shown graphically. Focus is on size of effect/pre-post.
Few-key text entry revisited: Mnemonic Gestures on Four Keyes	(Wobbrock, Myers, & Rothrock, 2006)	Input-device experiment (text entry)							Lab	10 sessions	RM-ANOVA to compare different input techniques over time with focus on main effect. However als analysis, whether devices were learnt faster with one technique compared to the other (pairwise comparisons).
Graffiti vs. Unistrokes: An Empirical Comparison	(Castellucci & MacKenzie, 2008)	Input device experiment							Lab	20 sessions	The study compares two input device techniques for typing on a PDA. A learning function for both devices is graphically displayed, statistical analysis focuses on pre-post comparisons.
How do different types of intragroup conflict affect group potency in virtual compared with face-to-face teams? A longitudinal study	(Lira, Ripoll, Peiro, & V., 2008)	Lab-based experiment							Lab	One month	4 sessions, but only 2 were used to gather data (first, last). Interesting analysis approach, but description lacks detail. Analysis uses a “moderated regression analyses” MRA framework (Cohen and Cohen 1983). Goal is to find predictors of change. Two predictors are within-subjects (conflict type) and one is between subjects (communication media).
Improved Search Engines and Navigation Preference in Personal Information Management	(Bergman, Beyth-Marom, Nachmias, Gradovitch, & Whittaker, 2008)	Survey							Field	Ca. 3 weeks (retrospective!)	A retrospective questionnaire that aimed at analyzing whether the use of a desktop search engine change the PIM behavior. A comparison group was tested as well.
Investigating Behavioral Variability in Web Search	(White & Drucker, Investigating Behavioral Variability in Web Search, 2007)	Log-based study							Field	5 months	The study analyses log fiels and individual differences in search behavior. Although it identifies patterns of usage, it does not analyze how these patterns might change over time. Basically, it tries to identify different user groups by identifying different patterns.

Title of the paper	Reference	Type of Study and data-gathering methods (e.g. experiment, ethnographic study, interviews, etc.)	**RQ: Average**	**RQ: Outcome**	**RQ: Pre-Post Comparisons**	**RQ: Shape**	**RQ: Event Occurence**	**RQ: In-depth**	Lab or field	Duration	Notes
It feels better than filing: Everyday Work Experiences in an Activity-Based Computing System	(Voida & Mynatt, 2009)	Logging, mid-term interviews, , final interviews after two months							Field	22-82 days	Very descriptive. Strangely they did not analyze changes, although they show a graph of log files over time. Instead only descriptive averages, graphical illustration of shape of change, but not commented nor analyzed. Qualitative description of people behavior (based on interviews) – only little references to changes, no systematic analysis here.
Let your users do the testing: a comparison of three remote asynchronous usability testing methods	(Bruun, Gull, Hofmeister, & Stage, 2009)	Longitudinal diary study							Field	5 days	Typical longitudinal study for interest in the average and NO interest at all in change processes. Quantitative comparison of number of outcomes (usability issues) with other testing conditions (e.g. experiment).
LiveRAC: Interactive Visual Exploration of System Management Time-Series Data	(McLachlan, Munzner, Koutsofios, & North, 2008)	Informal field study including interviews, notes, audio, desktop sharing and log data							Field	No information	Typical “longitudinal” study for interest in the average person and not so much in change processes. Qualitative summary of key usability issues.
Living with a Tabletop: Analysis and Observations of Long Term Office Use of a Multi-Touch Table	(Wigdor, Penn, Ryall, Esenther, & Shen, 2007)	Interviews, log file analysis, and analysis of email composition							Field	13 months	Graphical analysis of logfiles (heat map) No time-based analysis – so simple aggregation of data over time.
Longitudinal Study of Changes in Blogs	(Bogen II, et al., 2007)	Log study							field	2.5 months	The study uses different algorithms to investigate changes in blogs over time. Thereby, it focuses on graphical analysis and identifies a certain pattern for changes, namely differences between weekdays and weekens.
Longitudinal Study of Continuous Non-Speech Operated Mouse Pointer	(Sporka, Kurniawan, Mahmud, & Slavik, 2007)	Input device experiment + interview							Lab	5 days	The study uses Helmert contrast analysis to compare the performance metric over time. Additionally, subjective ratings are presented graphically over time.

Title of the paper	Reference	Type of Study and data-gathering methods (e.g. experiment, ethnographic study, interviews, etc.)	**RQ: Average**	**RQ: Outcome**	**RQ: Pre-Post Comparisons**	**RQ: Shape**	**RQ: Event Occurence**	**RQ: In-depth**	Lab or field	Duration	Notes
Longitudinal Study of People Learning to Use Continuous Voice-Based Cursor Control	(Harada, Wobbrock, Malkin, Bilmes, & Landay, 2009)	Fitts Law/Steering Law testing of input device, observation of use							Lab	2,5 weeks	Interesting study, as it includes a variety of longitudinal research question, although in a mostly explorative manner. It basically touches every kind of research question that is possible with focus on outcome and size of change. Although it was a lab-based study, it was highly explorative with few participants. Analysis: Quantitative: Purely descriptive statistics (percentage of improvement first-last session; performance at last session, shape as average among participants for each session). Qualitative: stories/episodes of usability issues, sometimes with comments whether they evolved over time, sometimes just as general observations – mostly on an individual level. No systematic analysis of changes.
Multimodal Capture of Consumer Intent in Retail	(Jain, Ghosh, & Dekhil, Multimodal Capture of Consumer Intent in Retail, 2008)	Diary study							Field	One week	Grounded Theory analysis of diary entries with goal of generating design guidelines. No time-based analysis.
Negotiatin Presence-in-Absence: Contact, Content, and Context	(Howard, Kjeldskov, Skov, Garnoes, & Grünberger, 2006)	Technology probe							Field	6 weeks	The study aims at understanding how presence in absence can be supported. Thereby, the study focused on qualitative analysis of episodes of usage and user comments. No emphasis on the analysis of changes was made.
Now Dasher! Dash Away! Longitudinal Study of Fast Text Entry by Eye Gaze.	(Tuisku, Majaranta, & Räihä, 2008)	Lab-based experiment							Lab	10 sessions	Problematic issue: time between sessions is different for each participant – still, session number is used as time variable. Analysis quantitaive: Mainly comparing first with last session. No statistical test reported. Graphical analysis of shape, no modeling of data and no comparison or analysis of growth parameters
On the passage of time: Temporal differences in video-mediated and face-to-face interaction	(van der Kleij, Paashuis, & Schraagen, 2005)	Paper-folding team work task in experiment							Lab	8 weeks (4 waves)	The study compares how people behave differently in a paper-folding collaborative task when either using face-to-face or video-mediated collaboration styles. RM-ANOVA was used to investigate differences over the 4 sessions. Graphical analysis was used to examine these differences
Privacy Diffusion on the Web: A Longitudinal Perspective	(Krishnamurthy & Wills, 2009)	Log analysis of 1200 popular websites and the additional websites visited							field	4 years (5 distinct data gathering waves)	The study presents extensive graphical analysis of data over time, interpreting the changes in terms of privacy diffusion in the web.

Title of the paper	Reference	Type of Study and data-gathering methods (e.g. experiment, ethnographic study, interviews, etc.)	**RQ: Average**	**RQ: Outcome**	**RQ: Pre-Post Comparisons**	**RQ: Shape**	**RQ: Event Occurence**	**RQ: In-depth**	Lab or field	Duration	Notes
Project Massive: Self-Regulation and Problematic Use of Online Gaming	(Seay & Kraut, 2007)	Survey Data							Field	14 months (3 waves)	The study analyzes over time, how online gaming is correlated with certain problematic behavior. It presents some aggregated data over time as well as a longitudinal regression model to investigate in changes over time in depth.
Robots in the wild: Understanding long-term use	(Sung, Christensen, & Grinter, 2009)	Observation and interviews							Field	6 months	Discusses very well the practical challenges of longitudinal research (participants not behaving as expected, not doing what they should do, not providing information the way it is planned). Good example for ad-hoc changes in a longitudinal qualitative study! Qualitative description of changes over time
Speech Technology in Real World Environment: Early Results from a Long Term Study	(Feng, Zhu, Hu, & Sears, 2008)	Logs, diaries and interviews							Field	6 months	The study compares two different user groups over the 6 months. Thereby, it aggregates data over time and uses the different groups as “change” variable (repeated cross-sectional design). The goal was to get more information about how people use speech technologies when interacting with computers.
Studying mobile context-aware social services in the wild	(Holleis, Wagner, Böhm, & Koolwaaij, 2010)	Field study with logging, questionnaire and final interviews							Field	4 weeks	Typical longitudinal “interest in average” study. Mainly: descriptive average statistics of tool usage. Some tests for changes over time (pre-post), but nothing in particular and no hypotheses.
Text Entry Performance of State of the Art Unconstrained Handwriting Recognition: A Longitudinal User Study	(Kirstensson & Denby, 2009)	Experiment and questionnaire							Lab	10 sessions	Interesting study. RM Anova to compare two different systems (within-subjects factor). Descriptive statistics for first and last session information. Graphical representation and discussion of change shape, including confidence intervals. Problem with RM-ANOVA: the different exposure times cannot be modeled (although not so important here, one can assume that it does not matter much). It is a bit unclear whether RM ANOVA was used to compare the overall means in performance and error or whether time was modeled.
The Design and Evaluation of a High-Performance Soft Keyboard	(MacKenzie & Zhang, The Design and Evaluation of a High-Performance Soft Keyboard, 1999)	Input device experiment							Lab	20 Sessions	The paper presents a model for keyboard text entry and tests this model with two different keyboard layouts. It compares these layouts pre-post and also presents graphical analysis of change process by indicating the “cross-over” point when the alternative keyboard outperformed the QWERTY layout.
Trends in Metadata Practices: A Longitudinal Study of Collection Federation	(Palmer, Zavalina, & Mustafoff, 2007)	Survey study (2 survey phases							Field	4 years (2 data-gathering waves)	Two-wave study which focuses on analyzing changes between the two data-gathering waves. Extensive information about mortality rate, response rate and type of participants. Descriptive analysis (no statistical tests). In the end, the data is cross-checked with additional interview data.

Title of the paper	Reference	Type of Study and data-gathering methods (e.g. experiment, ethnographic study, interviews, etc.)	**RQ: Average**	**RQ: Outcome**	**RQ: Pre-Post Comparisons**	**RQ: Shape**	**RQ: Event Occurence**	**RQ: In-depth**	Lab or field	Duration	Notes
Usability over Time	(Mendoza & Novick, 2005)	See section 2.2.4.1 Interest in the Shape of Change for a detailed discussion							field	8 weeks	See section 2.2.4.1 Interest in the Shape of Change for a detailed discussion
Veränderung in der Wahrnehmung und Bewertung interaktiver Produkte	(von Wilamowitz-Moellendorf, Hassenzahl, & Platz, 2007)	COPRPUS Interview-technique (developed in the paper)							Lab	14-33 months (retrospective!)	A retrospective interview technique which aims at analyzing how people change their judgment over time for a mobile phone, a Siemens medical system and a MS software (Powerpoint & Excel). Graphical analysis via tables to define the direction of change processes
Why structure and genre matter for users of digital information: A longitudinal experiment with readers of a web-based newspaper	(Vaughan & Dillon, 2006)	Lab-based experiment							Lab	5 Sessions	The study presents an extensive experiment of different web-based news readers to examine the influence of structure and genre. Different measures were taken and the data was analyzed over time both graphically and statistically (RM-ANOVA).