Humboldt-Universität zu Berlin

Institut für Bibliotheks- und Informationswissenschaft

Dissertation

Does it matter where we test?
Online user studies in digital libraries in natural environments

zur Erlangung des akademischen Grades

Doctor philosophiae (Dr. phil.)

Philosophische Fakultät I

Elke Greifeneder

Dekan: Prof. Michael Seadle, Ph.D.

Gutachter/in: 1. Prof. Michael Seadle, Ph.D.

2. Prof. Jeffrey Pomerantz, Ph.D

3. Prof. Pia Borlund, Ph.D

Datum der Einreichung: 10.04.2012 / Datum der Promotion: 04.06.2012

Abstract

Does it matter where we test? Online user studies

in digital libraries in natural environments

by Elke Greifeneder

User studies in digital libraries face two fundamental challenges. The first is the necessity of running more user studies in an online environment. Users can access digital library collections and services worldwide and the services should be usable at any time, because users access these services at a time and place of their choice. Online studies enable researchers to be separated from their participants in space (synchronous tests) and/or in time (asynchronous tests). This need for more online studies is coupled with a second need, a demand to test under realistic conditions outside of laboratories in users' natural environment.

Asynchronous remote usability tests are a methodological approach that might answer both needs: they allow participants to take part in a study at a time and place of their choice, often in the participants' natural environment. Any chosen place, however, might be noisy. Distractions are ubiquitous in a user's natural environment. An awareness of the potential influences of distractions on users' behavior during test situations is of great importance, because the validity of a study depends on the quality of the data. If an instrument allows systematic mistakes in measurements because of distractions, the validity is at risk. This dissertation examined if distraction in the users' natural environment produces a systematic mistake in digital library studies that take place at a time and location of participants' choice.

In order to investigate the existence of distractions during online user studies in digital libraries and to analyze the influence(s) of that distraction, a psychological experiment was set up. It examined completion time scores between participants in a laboratory (N = 38) and participants in their natural environment (N = 37). Both groups completed the same asynchronous remote usability test, which consisted of five search tasks in four digital libraries and in an online shop serving as control site. Survey data on the participants' distraction level during the test were collected.

The results of the experiment showed that participants were highly distracted and that participants in their natural environment needed more time to complete the same test. The setting did not affect successful task completions, the participants' judgments of sites or their decision-making processes. Multi-tasking, which seemed the obvious influencing distraction in the natural environment, did not alter the time scores in a significant way. Being contacted during the test, on the other side, changed the data in a significant way.

Based on these experimental findings, this dissertation developed a conceptual framework for online user studies in natural environments that suggests three types of variables that need to be collected: *core* variables that are necessary for data collection, *informative* variables that can help to interpret individual users' behavior, and *additional* variables that are not required but still can be useful for particular research questions. This work can conclude that it does not matter where we test, but it matters what happens during the test. The danger of data collection in a natural environment is not that events might occur, but that researchers know nothing about them.

Zusammenfassung

Does it matter where we test? Online user studies
in digital libraries in natural environments

von Elke Greifeneder

Die Benutzerforschung zu digitalen Bibliotheken sieht sich aktuell zwei großen Herausforderungen gegenüber: dem Bestreben, Studien vermehrt über das Internet durchzuführen, und dem Wunsch, Benutzerverhalten in natürlichen Umgebungen statt in Laborsituationen zu erforschen.

Benutzer greifen von ganz unterschiedlichen Orten auf der Welt auf digitale Angebote zu; daher liegt es nahe, digitale Bibliotheken über das Internet zu evaluieren. Solche Online-Studien erlauben eine räumliche Distanz zwischen Forscher und Teilnehmer; asynchrone Online-Studien – im Gegensatz zu synchronen – ermöglichen dabei zusätzlich eine zeitliche Distanz. Die methodische Herausforderung, Studien valide online durchführen zu können, ist mit der Herausforderung verbunden, Benutzerstudien unter realistischen Bedingungen, also außerhalb von Laboren, durchzuführen. Asynchrone Remote-Tests, hier am Beispiel eines asynchronen Remote-Usability-Tests durchgeführt, sind eine methodische Herangehensweise, die möglicherweise die Lösung beider Bedürfnisse sind. Sie erlauben Personen die Teilnahme an einer Studie zu einem Zeitpunkt und an einem Ort ihrer Wahl; der Ort der Wahl entspricht dabei in der Regel der natürlichen Nutzungsumgebung der Teilnehmer.

Doch in der natürlichen Umgebung sind Ablenkungen ubiquitär. Da die Validität einer Studie von der Qualität der Daten und deren Interpretierbarkeit abhängt, ist es für die Forschung sehr wichtig, die möglichen Einflüsse von Ablenkungen auf das Nutzerverhalten in der Testsituation mit zu bedenken. Wenn ein Messverfahren aufgrund von störenden Ablenkungsfaktoren bei einer Studie systematische Fehler produziert, ist die Validität der Studie in Gefahr. Das Dissertationsprojekt untersuchte, inwieweit asynchrone Remote-Studien solch einen systematischen Fehler bei der Evaluierung digitaler Bibliotheken produzieren.

In einem psychologischen Experiment wurde einerseits das Vorhandensein von Ablenkung während der Testdurchführung in natürlichen Umgebungen ermittelt und andererseits der Einfluss dieser Ablenkung auf das Nutzungsverhalten analysiert. Experimentell wurde die Zeit gemessen, die

Teilnehmer in einem Labor (N = 38) und Teilnehmer in ihrer natürlichen Umgebung (N = 37) zur Fertigstellung des Tests benötigten. Beide Gruppen absolvierten denselben asynchronen Remote-Usability-Test, der aus fünf Suchaufgaben in vier digitalen Bibliotheken und einem Online-Shop bestand. Ein Fragebogen erfasste zusätzliche Informationen über die Art der Ablenkung der Teilnehmer.

Die Ergebnisse des Experiments zeigen, dass die Remote-Teilnehmer während der Studie stark abgelenkt waren und dass sie in ihrer natürlichen Umgebung deutlich mehr Zeit für denselben Test benötigten. Der Ort der Testdurchführung beeinträchtigte jedoch statistisch gesehen weder die Erfolgsquote bei der Erledigung der Aufgaben noch die abgegebenen Bewertungen der Studienteilnehmer noch ihren Entscheidungsprozess. Multitasking während des Tests – die augenfälligste Ablenkung in der natürlichen Umgebung – veränderte die Durchführungszeiten kaum. Wurde ein Teilnehmer jedoch während des Tests aktiv kontaktiert, führte dies zu einem statistisch signifikanten Unterschied zwischen den beiden Testgruppen.

Basierend auf diesen experimentellen Ergebnissen präsentiert die Dissertation ein konzeptuelles Framework für Online-Benutzerstudien in natürlichen Umgebungen. Das Modell schlägt drei Arten von Variablen vor, die für die Auswertung der Daten aus diesen Studien benötigt werden. Diese sind Kernvariablen, informative Variablen (die bei der Interpretation von Verhaltensmustern individueller Teilnehmer helfen können) und zusätzliche Variablen (die nur für bestimmte Forschungsfragen relevant sind).

Aus den Ergebnissen des Dissertationsprojekts folgt, dass der Ort der Testdurchführung nicht relevant ist, aber dass es von großer Bedeutung für die Validität und Interpretierbarkeit der Daten ist, im Test zu erheben, was während der Durchführung in der natürlichen Umgebung des Teilnehmers geschieht.

Acknowledgments

A friend of mine told me once that writing a dissertation is nothing more than writing a longer master thesis—she was utterly wrong. A dissertation is so much more. It makes you realize what the word *research* really means. There are a number of people who have guided me along this learning process and I would like to thank them for their support.

First of all I most like to thank my Doktorvater Michael Seadle, who guided me through these years of research, both intellectually and socially. He has offered me ongoing support, inspiration and encouragement to continue my path in the research area of my choice. Not enough thanks can be given to the members of my thesis committee—Jeffrey Pomerantz and Pia Borlund—who moved across oceans and beyond languages and offered support in so many ways. I would like to extend these thanks to my additional dissertation defense committee members, Stefan Gradmann and Frank Havemann. I would also like to thank numerous researchers for their insightful and challenging comments on my dissertation during various occasions, in particular to the participants of our doctoral colloquium.

This research could not have been carried out without a number of students who spent endless smiles to recruit the ideal sample for this study. This warm thank you is for Maria Yalpani, Pamela Aust, Ulrike Stöckel, Katja Metz, Nadine Messerschmidt, Kristin Reinhardt and Lars Gottschalk. I also thank the students of the Master course who participated in the pilot test and all anonymous participants of the final experiment.

I would like to thank the software producers of *Loop11* for the free use of their product and for ongoing support by email, as well as the university library for allowing me to recruit participants in their new library building.

A dissertation hopefully ends on the up-side of a research project, but there were doubtless many downs as well. Many friends have helped me, and offered the necessary stability from getting crazy. I would like to thank in particular Vivien Petras, Leah Rosenblum, Sandra Lechelt and the one person in my life who guided me with his love and made the darkest places still have a touch of sunlight. I am glad he exists: thank you Thomas for everything.

This dissertation would not have been possible without my family and I most gratefully thank my two brothers Rainer and Jürgen, who helped me each in their own way to do the right things. Above all, I would like to thank my parents, who made this all possible with their ongoing support for so many years, their encouragements and their love.

This thesis is dedicated to you.

Table of Contents

List of Tables

List of Figures

Abbreviations

Amazon	German website of the online shop “Amazon”
Bundesarchiv	Digital library “Digital Picture Archives of the Federal Archives”
DigiZeitschriften	Digital library “DigiZeitschriften”
LAB	Laboratory
NE	Natural environment
ORKA	Digital library “Open Repository Kassel”
Perseus	Digital library “Perseus Digital Library”
SSOAR	Digital library “Social Science Open Access Repository”
Valley of the Shadow	Digital library “Valley of the Shadow”

1 Introduction

1.1 Problem statement

The secret of a good dish is high quality ingredients. The same is true for user studies in digital libraries: the validity of a user study depends on the quality of the data. Validity describes whether an instrument actually measures what was intended to be measured. The fewer systematic mistakes an instrument makes, the higher the validity of a study. This dissertation examined if distraction in the users' natural environment produces a systematic mistake in digital library studies that take place at a time and location of participants' choice.

People interact with information in a noisy world: distraction is part of our daily life. Mobile computing makes people perpetually reachable—the mobile phone with internet connection to the world is always turned on. People use information systems in public spaces and at home, i.e in an environment in which the simultaneous use of several devices has become the rule and not the exception. The public and private space in which everyday information interaction takes place is called the natural environment. The natural environment is the opposite of a laboratory. It is any place outside the laboratory where users choose to take part in the test.

The noisiness of the natural environment has several implications for information behavior studies and for digital library user studies in particular. Arms (2000) defined a digital library as a "managed collection of information, with associated services". Users can access digital library collections and services worldwide and the services should be usable at any time, because users access these services at a time and place of their choice. Users come from different organizations and from different cultures; they speak different languages and live in different time-zones. Studying digital library users is a real challenge.

Digital services are a central part of libraries and their user-orientation is indispensable. However, it is not always clear how to create digital libraries that actually match users' needs. Despite a multitude of golden rules (Shneiderman & Pleasant, 2005), advice (Krug & Dubau, 2006) or guidelines (e.g. Nielsen, 2000), there is no magic formula that works for all users and all services. A user-oriented digital library service can only be achieved by recurring (iterative) studies about and with users.

User studies are not restricted to one area of library and information science; instead they are applied in information-seeking behavior studies, in relevance measurements, in usability studies, in user experience design, and in many more areas. Their common element is a cognitive user-centered approach: there is a genuine need to understand users and their context (Seadle, 2000 and Sexton et al., 2004). While they share the same understanding and frequently the same methods, their perspectives on the outcome of user studies vary. For example, information retrieval researchers are interested in search behavior, whereas digital library designers aim at usable products.

1.2 Research background

The need to understand digital library users and their context is not a new phenomenon. Pomerantz et al. (2008) stated in their work on the *Development and Impact of Digital Library Funding in the United States* that the *Digital Library Initiative* of the 90s paid “too little attention to user needs, and too little attention to evaluation” (Pomerantz et al., 2008, p. 49). Recently, this situation has changed with a growing attention to user needs and many evaluations have been carried out (Greifeneder, 2010).

While some notable exceptions collected high quality data, many of these studies fail to demonstrate validity or show a deliberate choice for good-enough-data. In 2002, Troll Covey interviewed participants from the Digital Library Federation (DLF) about their use of and experience with methods in digital library user studies and stated:

> “Libraries are struggling to find the right measures on which to base their decisions. DLF respondents expressed concern that data are being gathered for historical reasons or because they are easy to gather, rather than because they serve useful, articulated purposes.” (Troll Covey, 2002, p. 2–3)

Nearly ten years later, Fagan (2010) and Lyons (2011) reinforced that statement:

> “In general, these studies [by librarians] used fewer participants […], followed less rigorous methods, and were not subjected to statistical tests.” (Fagan, 2010, p. 61)

> “Like the most impatient of information seekers, we ignore the fact that inadequate information gathering techniques will lead us quite expediently to the wrong answers. Neither do our national and international library organizations set consistently good examples in this regard. Too often they employ deficient research methods or promote unjustifiable interpretations of data they have collected.” (Lyons, 2011, p. 92)

Bargas-Avila & Hornbaek (2011) reviewed empirical research on user experience testing and examined publications from 2005–2009. They discovered that more "than half of the publications used questionnaires as a way of assessing UX […]. Half of them (51%) use self-developed questionnaires but do not provide readers with the items used." (Bargas-Avila & Hornbaek, 2011, p. 575). Julien et al. (2011) reported that surveys in information behavior research are "still the largest proportion of methods used" (p. 21). A similar content analysis with a tighter focus on digital library user studies from 1998–2008 found that 43% of user studies used surveys and 18% used log file analyses. Most of these studies, including the ones using surveys, examined the use of digital libraries (Greifeneder, 2010).

Are surveys really the best method for digital library user studies? Few librarians are trained to be ethnographers or social scientists and surveys appear to be an easy enough data collection method to adapt to local purposes (even if survey development is nothing but easy). Until a few years ago, surveys or log file analyses were the only two methods that allowed most researchers to collect data online, that is, by means of the internet. As a result, user studies on digital libraries were mostly surveys (Xia, 2003; González-Teruel et al., 2004; Liu, 2006; Bayram & Doğan, 2006; IRN Research, 2011) or log file analyses (Bogros, 2003; Nicholas et al., 2006; Safley, 2006; Feran et al., 2007). Surveys and log file analyses were used because of convenience (they were easily implemented), because of historical reasons (they were the two predominant methods that allowed a collection of online data), and because they were at that time the best methods (when used appropriately, they can produce valuable data).

Laboratory studies are the prevalent alternative to surveys or log file analyses for digital library user studies. Especially in the area of information retrieval, laboratories are the established setting. The refereed journal *Online Information Review* published a special issue on *Evaluating web search engines (vol. 35, n°6)* in late 2011. All of the studies either excluded users completely or carried out the research in a laboratory.

The decision to carry out studies in a laboratory or at least "offline" can be problematic. "Offline" means that participants must be available where the research takes place ("offline" does not necessarily imply an artificial place like a laboratory, for example it could be a focus group). Depending on the research question, running a digital library user study "offline" can reduce the possible user population to a minimum. In a laboratory, participants are placed in an unfamiliar setting in which almost all context is predefined by the researcher. Results from laboratories do not take into account that participants might use different operating systems, different system configurations, different internet connections or a different screen resolution and might therefore show a different behavior. In laboratories, distraction is eliminated as a confounding variable and is not treated as an influential part of the users' information environment.

With few exceptions, digital libraries are accessible online. This does not imply that all user studies on digital libraries should now be undertaken online. There will probably always be specific research questions which can be better answered in a laboratory or "offline" setting. However, online studies have one major advantage: the research takes place where the users are.

The call for more online studies is widespread across disciplines: the most frequent arguments refer to reduced traveling costs and the usage of distributed sampling techniques (Thompson, 2004; Gardner, 2007; Huang, 2009). These arguments are made in particular for studying the development of Open Source Software (Andreasen, 2007), hard-to-reach populations like blind users (Mankoff, 2005) or in order to draw a better picture of culturally diverse users (Baker, 2007; Lee & Lee, 2007; Clemmensen, 2007). And it is not only the fact of being able to test online that intrigues researchers. It is also the possibility of larger scale testing. Field-studies of online spaces like *YouTube*, *Flickr* or *Facebook* allow studies at large scales, but require methods that can be applied online (Rotman et al., 2012).

But the two predominant methods for online studies—surveys and log file analyses—both have handicaps. Surveys collect claimed behavior: that is, they only report what participants said what they think they did. This can match reality or it can be a false impression. Log file analyses, on the other hand, report actual behavior—how users actually interacted with the system—but they do not tell researchers the reasons behind that behavior. Even if a large number of usability surveys are available (for example *Attrakdiff2*, *IsoMetrics*, *SUMI*, *QUIS*), usability practitioners long preferred "thinking aloud usability testing" in a laboratory. This method asks users to complete concrete tasks and to comment loudly on one's own behavior. In contrast to surveys or log file analyses, "thinking aloud usability testing" allows researchers to compare the results of specific task completion between participants, and allows them to enrich a quantitative comparison of expected and actual behavior with qualitative data.

This put usability researchers in a quandary: they wanted to run online tests (because they produced products for an international audience), but they lacked appropriate methods to run traditional usability tests online. Out of that need, synchronous and asynchronous remote usability tests were born. “Remote” in this case means that the test takes place at a geographical distance between researcher and participant. “Synchronous” implies that there is no additional temporal distance, while “asynchronous” adds a temporal distance to the test setting. A survey is in this sense an asynchronous remote test, because the participants can take part in the test at a place and time of their choice. Both approaches—synchronous and asynchronous remote usability tests—will be discussed more in-depth in chapter 2.

This dissertation focuses on asynchronous remote usability tests, which can be adapted as widely as the various survey implementations. The method offers an online setting in which researchers can examine predefined task completion. Asynchronous remote usability tests even allow online experimental settings, which were previously restricted to local settings, i.e. mostly laboratories. In addition, asynchronous remote usability tests are easy to set up technically and are comparatively cheap.

It is no surprise that the marketers jumped on this research area and have recently dominated it. Online marketing research has been a prosperous area since its beginning (Gold, 2008) and companies invested large sums in early studies to identify the possibilities and boundaries of remote tests like the early IBM study (Bartek & Cheatham, 2003) or the study by Boldt|Peters (2005).

Because asynchronous remote usability tests are easily set up, fairly cheap and badly needed, there is a danger that researchers will glorify the possibilities and underestimate the limits. Bustamante (2010) ran an asynchronous remote usability test to examine information seeking on a website and concluded that the software *Loop11* (which will be used in this dissertation as well) produced clear, objective, comparable and convincing data (“*Loop11* nos permite presentar los resultados de un test de manera clara, objetiva y convincente“; “Estos datos objetivos y comparables” (Bustamante, 2010, p. 428 and p. 429)).

Bustamante's statement brings up the question from the beginning of this dissertation: do asynchronous remote usability tests measure what they intend to measure and produce data that are clear, objective and comparable? Or do the numbers only look objective and comparable, when in fact the stories behind the numbers are very different?

Running an online user study not only means that participants are somewhere in the internet, but that the participants take part by means of the internet, that is, they are still in front of a machine in their own information-use context. And this context is noisy.

While library and information scholars have historically played a marginal role in the field of online studies, a large quantity of research on users' context exists. Calls for context research have increased in recent years such as that from Snow et al. (2008) who demanded an "understanding of the issues surrounding usage of digital objects" (Snow, online) and Bargas-Avila & Hornbaek (2011) who criticized that the "[c]ontext of use and anticipated use, often named key factors of UX, are rarely researched" (p. 575) and that most online studies do not describe the use context at all. In a highly cited keynote at the *Information Interaction in Context* conference 2010, Saracevic developed five axioms on context. His second axiom stated that "[c]ontext is not self-revealing, nor is it self-evident. Context may be difficult to formulate and synthesize. But plenty can go wrong when not taken into consideration in interactions" (Saracevic, 2010, p. 1).

Assuming that their usage context is noisy, participants are prone to be distracted in their information use behavior. But if distraction is inseparably bound to the natural environment and the natural environment is the place in which asynchronous remote usability tests take place, then knowing about the influence of distraction is as important for data interpretation as is the knowledge that surveys produce data solely on claimed behavior. Or to paraphrase Saracevic: plenty can go wrong when *distraction* is not taken into consideration.

There have been a number of studies that have investigated users' context. *PooDLE* was a project at Rutgers University School of Communication and Information under the direction of Nicholas Belkin (Liu et al., 2010). It built a "personalization assistant for personalized interaction with digital libraries" (Rutgers University School of Communication and Information, 2009, online). The project description on the website stated that "there has to date been little research done on how to unobtrusively discover relevant characteristics of a searcher and the searcher's context". The project group conducted a controlled experiment in a laboratory in which participants had to complete four tasks and had to fill out a mostly quantitative survey afterwards. The notion of "searcher's context" was limited to participant's individual cognitive space.

The *Web Use Project* at Princeton University conducted a study on users' search activities and collected data on participants "online skills in the context of their social attributes" (Hargittai, 2002, p. 1239). What they actually collected were survey data on "the frequency and location of respondents' regular Internet use, the types of sites they visit, the types of activities they perform online, their use of other media, their time spent on various social activities and their social support networks" (Hargittai, p. 1241). This form of context describes participants' information use behavior within the sphere of the internet.

Griffiths & King (2007) studied "the relationships between physical spaces, such as museums, their visitors and physical and virtual visits—on-line users and uses" and conducted a national survey "of the information needs and expectations of users and potential users of on-line information" (Griffiths & King, 2007, online). Context of use, as it is understood by Griffiths & King, meant communication mechanisms, resources and contents. They collected information on the frequency of internet use, providers and access mode. This understanding of context is still participant-oriented and excludes the influence of the home or the work environment on use behavior.

Rieh (2004) used a qualitative approach to gather data on information-seeking in home environments. She commented that "a shift in Internet use from work to home involves far more complex factors than physical setting alone, because home provides social context for diverse information activities including seeking, use, and evaluation" (Rieh, 2004, p. 743). While she provided valuable information about the home environment—such as where the computer was installed—, she left out further details about the "complex factors" a home environment offers. Rieh did not mention distraction as a factor.

Kelly (2006a and 2006b) used a "naturalistic" approach to collect data "about information seeking context and behavior in natural information seeking environments [in order] to identify which aspects of context should be considered when studying information seeking" (Kelly, 2006a, p. 1730). She installed tracking software on participants' machines and examined their information-seeking behavior in their own environment. While her understanding of context included the participants' personal machines and settings, her notion of context stopped at the edge of the computer. Her study described information-seeking behavior in a natural environment, while this research describes the natural environment, in which information-seeking takes place.

1.3 Research questions

Researchers need to understand how a particular research design might alter results. Many researchers conduct user studies in digital libraries, but only a few examine the way research is conducted and the way research designs affect results. The introductory literature review has shown that some researchers even complain about a general lack of data validity in digital library user studies.

There are, of course, a few exceptions in the library and information science community who examined effects on research designs. Borlund (2000) discussed various choices of experimental components for interactive information retrieval tests and Kittur (2008) investigated the utility of a micro-task market for collecting user measurements. Kelly et al. (2008) examined the influence of questionnaire modes on subjects' usability ratings and on responses to open questions. In addition, more than 30 studies compared laboratory and remote (both synchronous and asynchronous) test settings. Some of these will be presented in chapter 2. This study has a different focus and compares data from a laboratory and from a natural environment (and not laboratory and remote).

Laboratories have long been seen as reasonable proxies for user studies. Because of the absence of confounding variables, laboratories allow researchers to work in a controlled environment and to assign a particular phenomenon to a single concrete behavior, but worldwide access to digital libraries has led to a greater variety of users with a broad spread of cultural backgrounds. Bringing a representative sample of this group to the laboratory could be a costly challenge because of travel expenses alone.

The literature review above has shown that there is a clear need for methods that allow running user studies in an online environment and that most studies respond to this need for online testing by conducting surveys or log file analyses. This dissertation focuses on an approach called asynchronous remote usability tests, in which researchers and participants are separated by time and space.

Participants have no direct contact with the researcher and can access the test wherever and whenever they like.

There is also a demonstrated need for more information about the context in which digital library use and the information-seeking in particular takes place. While there are many studies on context, earlier research has focused on individual cognitive spaces or at most on users' own machines. No work has explicitly studied the impact of distraction on users' behavior.

This means that the need for online studies is coupled with a demand to test under realistic conditions—away from laboratories. These conditions grow out of the users' natural environment where people simultaneously use a digital library, join a chat or read an incoming *Facebook* post. The effects of these disruptions generate a gap that is generally not taken into account. This dissertation seeks to close that gap and examines to what degree distraction plays a role and affects data from asynchronous remote usability tests.

Without knowledge about possible influences on the data, the numbers can have little meaning. If a participant needed 400 seconds to complete a task in a user study, it can mean that this person actually needed 400 seconds, because the website had poor usability or because the participant had "bad luck" or got "lost" on the site (Nielsen, 2006). However, in a natural environment, 400 seconds on a task can also mean that a participant only needed 50 seconds to complete the task and read an incoming email during the remaining 350 seconds before signaling that the task was completed.

A clear understanding of the possibilities and limits of this methodological approach are of great importance, because asynchronous remote usability tests can be applied in most domains of library and information science: they can be used for usability testing or for information-seeking behavior; they allow naturalistic information behavior studies (for example on user satisficing), and for the first time they allow researchers to test digital libraries at a distance on users' own devices—from the notebook to the eBook Reader as well as to run experiments outside of laboratories.

On the basis of these research needs, this dissertation asks whether the natural environment with its distractions has an impact on digital library user studies. Or to put it more plainly: Does it matter where we test? While this question seems at a first glance to be rhetorical, researchers lack evidence for a definite answer. Laboratories may have been reasonable proxies for user studies for many decades, but it is unclear if this holds true in today's noisy natural environment. If it matters where we test, that is if the data is different between a laboratory study and the same study in a participant's natural environment, then the researchers need to know how much the data has been altered by the test setting. If the test setting alters the data, researchers need to collect additional variables in user studies to enable researchers to make an adequate interpretation of the data. Therefore, this dissertation's goals are to

- examine if users are distracted in their natural environment while they participate in an asynchronous remote usability test, and to examine how much this distraction influences test data;
- develop a framework of *core*, *informative* and *additional* variables that enables researchers to collect and interpret data in asynchronous remote usability test settings.

This dissertation examines information behavior in the sense of how the natural environment as a user test setting influences how human beings interact with information (Bates, 2010). This information interaction can be positive or negative. Because of the influence of the natural environment users may also ignore, deny or reject information (Case et al., 2005). This research does not seek to examine (interactive) information retrieval in the sense of how the natural environment might change information searching and might lead to different search tactics.

The two principal research questions for this dissertation are:

(RQ 1) Are there differences in the data gathered from the same test in a laboratory and in a participant's natural environment?

(RQ 2) Is distraction the cause of that difference, if it exists?

1.4 Structure of dissertation

A psychological experiment was designed in which participants had to complete brief search tasks. One group of participants completed an asynchronous remote usability test in a laboratory and another group completed the same test in their own natural environment. The experiment was conducted in May 2011 at the Humboldt-Universität zu Berlin.

The dissertation uses standard statistical techniques to compare data from the two settings—laboratory and natural environment—and uses an additional description of outliers' behavior to illustrate individual users' behavior. The focus of this research is on the measurement of time in asynchronous remote usability tests in a natural environment and on an exploration of necessary control variables. These two elements will be examined within the context of other elements in user tests such as task completion and judgments.

Chapter 2 explains online studies, in particular synchronous and asynchronous remote usability tests, and provides further information about the natural environment as well as the influence of distraction on behavior as documented in related research. Chapter 3 reports results from a pilot test of the experiment, and chapter 4 describes research procedures, including the method and recruitment process. Chapter 5 discusses outliers and provides a description of participants' individual behavior in an asynchronous remote usability test setting. A formal data description is provided in chapter 6. Chapters 7 and 8 present the findings on the two research questions with chapter seven examining the differences between the settings and chapter 8 analyzing the influence of control variables. The findings result in a conceptual framework for online user studies in natural environments, which will be discussed in chapter 9.

2 Theoretical framework

2.1 Online studies

2.1.1 Online methods

Online studies do not necessitate new methods, because online methods are by and large adaptations of traditional methods, but they require new thinking about research designs. Markham & Baym's 2009 statement that researchers were "naive enough to think that it would be relatively straightforward to transfer research strategies developed for studying face-to-face contexts to life online" (p. viii) described the most serious problem: the misconception that running online studies meant using the same data interpretation mechanisms—just in a new environment.

This chapter[1] gives an overview of online methods as well as application scenarios and potential risks. The experiment described in this dissertation draws on asynchronous remote usability tests, but the findings have implications for most of the methods described in this chapter, because most participants in online studies are in their natural environment.

The discussion about online methods started several years ago. The first online studies were undertaken in 1995 (Wenzel & Hofmann, 2005) and the refereed journal *Cyberpsychology* launched its first special issue on online methods back in 1999. After more than 15 years of online studies, researchers have slowly lost their enthusiastic and simplistic view of their possibilities and started to acknowledge the boundaries as well.

Online studies have a dual relationship with the internet: the internet is both methodological tool as well as object. In other words, researchers collect data about the internet by means of the internet (Welker & Wenzel, 2007; Orgad, 2009). Online studies are also known as "Internet Research" (Markham & Baym, 2009) or as "Virtual Research" (Buchanan, 2004; Hine, 2006).

The characteristic trait of online studies is a spatial distance between researcher and participant. This trait is labeled remote. The term "online" describes the test environment; the term "remote" describes the form of connection between researcher and participant. In a synchronous remote

[1] Parts of the chapter on online methods have been published in a book chapter in German in Greifeneder, E. (2011a). Einführung in die Online-Benutzerforschung zu Digitalen Bibliotheken. In: B. Bekavac, R. Schneider & W. Schweibenz (Eds.), *Benutzerorientierte Bibliotheken im Web* (pp.75–94). Berlin: De Gruyter Saur.

setting (also called a moderated remote test), researchers and participants are separated in space, but they have a real-time connection using text, voice or video. Its advantages over laboratory tests are the elimination of travel costs and the provision of a familiar environment for participants. Asynchronous tests (also called unmoderated tests) appeared more recently and add a temporal dimension. Researchers and participants are now separated in time as well as space. Participants have no direct contact with the researcher and can access the test at a place and time of their convenience.

The appeal of digital libraries also creates a key barrier for user studies: the lack of temporal and spatial constraints. Users are distributed around the world and it can be a challenge to persuade digital library users from China or the Unites States to come to a focus group in Germany or to run a synchronous chat-interview across different time zones.

Some methods in online studies use tools that have genuinely been developed for a specific online method (like the web survey or remote usability tools); other methods draw on tools that were designed for other purposes (like chat rooms or virtual video conference facilities).

For the last ten years, web surveys (also called online surveys) were the most popular method in online studies, because from a technical perspective their implementation is straightforward. Ready-to-go software solutions like *LimeSurvey* or *SurveyMonkey* make it easy to produce a survey and, thanks to good export functions, to receive edited data in diagrams and charts for immediate use in presentations. This software's technical simplicity makes it possible to develop a survey quickly and cheaply.

However, the technical simplicity of surveys conceals difficulties in question development and limits in data interpretation. Asking correct and unambiguous questions requires a high degree of problem awareness. In addition, surveys only capture claimed behavior and researchers never know to what degree users' claimed behavior matches their real behavior.

Log file analyses collect actual user behavior, for example about the time and form of a user's interaction with a system. The term originates from the navy where captains recorded their current travel position in a logbook. The method is appropriate for pattern discovery like referrer sites, peak times, or sites with a high number of error notices. Many studies use log file analyses to make statements about page views, downloads, or zero results. Log files can be used to visualize usage statistics. The most common approaches are mouse-tracking or click-tracking, which may be visualized in the form of heat maps.

The difficulty with log file analyses is hidden in the interpretation of the logs. Researchers struggle with how to define a visit to an online service, or how to define what a page view is on a site, not to mention the discussion what makes a digital library visit a success (Troll Covey, 2002). There is also the problem of crawlers which might be misinterpreted as human visits. And log file analyses have another disadvantage: they only report the kind of interaction that occurred and not the reason for it. Log file analyses cannot answer cause-relationship-questions.

The following three methods produce qualitative data and use externally available tools to contact users via the internet. These methods are the online interview, the online focus group and the online (or virtual) observation. The difficulties of qualitative online studies are evident: how can a researcher conduct a qualitative interview with someone who is neither in the same room nor in the same time zone? In order to take part in an oral online interview, participants need technical equipment for Voice-over-IP, like microphones or headsets.

Apart from different time zones, different user languages make qualitative interviews challenging. Users of a digital library like *Europeana.eu* speak many different languages. Researchers can run all interviews in English, for example, or they can offer the interview in as many different languages as possible. A reduction to English is disadvantageous, because it is likely that neither the interviewer nor the interviewees are native English speakers and misunderstandings are essentially pre-programmed. Even finding qualified researchers who speak many languages is challenging and too many interviewers result in interviewer chaos. Of course, the problem of language also exists for “offline” interviews, but it becomes more pressing with online, international environments.

Interview time is another problem, because researchers at the home institution may prefer not to conduct interviews at night so participants living in different time zones may be systematically excluded. Hence, online interviews risk producing a sampling bias. Qualitative interviews require a certain amount of time: a brief interview takes 30 minutes and an intensive interview can take up to 90 minutes. However, few online users are prepared to spontaneously spend an hour participating in an unscheduled interview.

In principle, two tools exist to run online interviews. The most common approaches are chat or Voice-over-IP interviews. The advantage of chat interviews is that the interview material already exists in written form. The disadvantage is that participants tend to say more than they write. There is also a risk of losing too many participants: a study by Stieger & Göritz (2006) examined the feasibility of instant messaging interviews and reported that 9.4% abandoned the interview prematurely. An example of an application of synchronous chat interviews as an online method was presented at the *iConference* 2011 (Bullard & O'Brien, 2011).

An asynchronous alternative takes the form of email interviews, which are characterized by a delay in the participant's response. Again, there is a potential risk of losing participants during the study, since many simply stop responding after the second or third email. Several studies discovered that email answers tend to be shorter than instant messaging, but that there is no difference in the meaning (Hussain & Griffiths, 2009; Meho, 2006). Kazmer & Xie (2008) offer a useful overview of techniques, problems and limits of online interviews.

Online focus groups operate similarly to online interviews with the difference that more participants attend and that the aim of an online focus group is different: an interview examines the opinions and the behavior of an individual, while focus groups are interested in a group result. The circumstances for online focus groups are similar to the ones described above on online interviews. Chase & Alvarez (2000) offer a useful introduction to the method.

The easiest (and most common) approaches are still asynchronous online focus groups using forums in which researchers start a question and participants can answer at a time and place of their choice. This form is rather difficult for a moderator, because silent participants are hard to encourage in such a forum where every comment will be published. An immature idea, mentioned in an oral discussion, is forgotten after a few minutes. The same idea in written form will be visible until the end of the focus group, which can turn into a barrier to participation. A good moderator can try to reduce these risks by preparing participants and proposing clear rules.

Online focus groups can be hosted in so called multi user dungeons or in virtual worlds like *Second life*. At the *American Library Association Annual Conference* in New Orleans, Haefele & Ray (2011) reported that they used *Adobe Connect* and *Wimba* to hold online focus groups. The future direction of online focus groups appears to be small virtual conference meeting rooms in which participants can “come” and join without additional software installations. The only requirements are a microphone and a headset, though a webcam is beneficial. The virtual rooms originate from the idea of small private chat rooms, which were frequently offered as an opportunity to retreat from a public chat. At present a wide range of products exist, but few are persuasive in terms of the quality of video and sound. Integrated whiteboards (for example *twiddla* or *flockdraw*) are an additional advantage offered by these virtual meeting rooms, because participants can draw or comment on them. Using online whiteboards even permits conducting “card-sorting tests” online, which is an effective way of comparing users’ mental models with the designers’ expectations.

Ethnographic observations are an invaluable part of user studies and the area is richly researched (for an overview see Miller & Slater, 2000). Hanging out in Geertz’s sense (Geertz, 2009) can be done unobtrusively in the internet—one simply goes online. The real challenge is concealed in form of the data that researchers can get in ethnographic observations on the internet. Online observation draws on text and pictures available on the internet. In that sense, online observation resembles very much a content analysis. An example of an ethnographic observation that goes beyond that is an ongoing project at North Carolina State University, in which researchers developed SUMA (Mobile Space Assessment Toolkit), i.e. an application for the iPad and other tablet PCs that allows virtual user tracking. The study is designed to give researchers a better understanding of the use of library spaces, and to help them to discover how virtual and real rooms are connected (Casden, 2011).

2.1.2 Synchronous and asynchronous remote usability tests

Usability is an essential part of user studies. Usability examines how effective and efficient a product is in use and how satisfied users are with the product (ISO 9241-210). Usability measures whether the structure and design of a digital service match users’ needs. While early usability studies focused solely on the effectiveness and efficiency of a product, usability design has since moved from merely usable products to good experiences with products.

The disadvantage of most usability methods are their dependence on a local test setting; that means that participants must come to a laboratory in order to participate in the test. In a laboratory, usability designers have their computers with mock-ups of the future website and frequently an eye tracker, which allows a detailed tracking of eye movement. A frequent usability method is “think-aloud-testing”, in which participants comment out loud on their behavior.

As explained earlier, usability designers were in a quandary as they looked at the limitations of user study techniques, because they valued their traditional approaches while at the same time seeing the need to test in online environments. A glance at the conference program of the German conference *Usability Professionals 2009* illustrates the initial reaction to the dilemma: an increase in the usage of standardized web-surveys (Beschnitt, 2009). But these surveys collected claimed behavior and lacked the possibility of testing task completion.

Since 2008, remote usability tests have gained in importance. At the *UXcamp Europe* 2010 in Berlin, which is the annual gathering place for European usability designers, remote usability tests were a recurring theme and the enthusiasm for the approach was immense. This form of testing allows users to participate without being restricted by spatial constraints. Remote usability tests are not a new method; they are a technical bridge to adapt “offline laboratory methods” to a remote setting.

Synchronous remote usability tests use common internet tools like screen sharing software to get into contact with participants. This technology allows researchers to access participants’ desktop. Chat or Voice-over-IP services enhance this experience. Researchers can ask questions and ask participants to complete tasks and perceive how a service or a site looks on the users’ own machines. They can follow participants as in usability think-aloud-tests. In contrast to a laboratory, they disturb participants less, because they are not peering over the participants’ shoulders. This mode of remote study was heavily used in early forms of remote usability testing.

The advantages of synchronous remote usability tests are that they offer a worldwide application spectrum at low costs, and that they allow an interaction with participants, including call-backs for clarifications. The temporal dependence—researchers must schedule appointments with participants—and the strong dependence on available technology on both the researchers’ and the participants’ sides exclude some potential participants. With synchronous remote usability tests, researchers run the risk of a biased sample. However, few usability think-aloud-tests are representative. Their aim is to detect problem areas and not to prove a research hypothesis. Representativeness is less important for most usability designers, but if researchers want to validate hypotheses with synchronous remote usability tests, lack of representativeness becomes an obstacle.

Researchers also work under restrictive legal and ethical conditions. For example, during a synchronous remote usability test, researchers gain full control over the participant's desktop and mouse: private information on the desktop as well as login data could be captured. It is highly recommended to share a written agreement with the participants about their rights and informing them about the researchers' ethical duties.

The desire to escape the problem of time constraints spurred researchers to come up with new solutions. Asynchronous remote usability tests add a temporal dimension: researchers and participants are now separated in time as well as place. Participants have no direct contact with the researcher and can access the test at their place and time of convenience. With asynchronous tests, bigger samples are possible and researchers can even run tests on mobile devices such as eBook readers or smartphones.

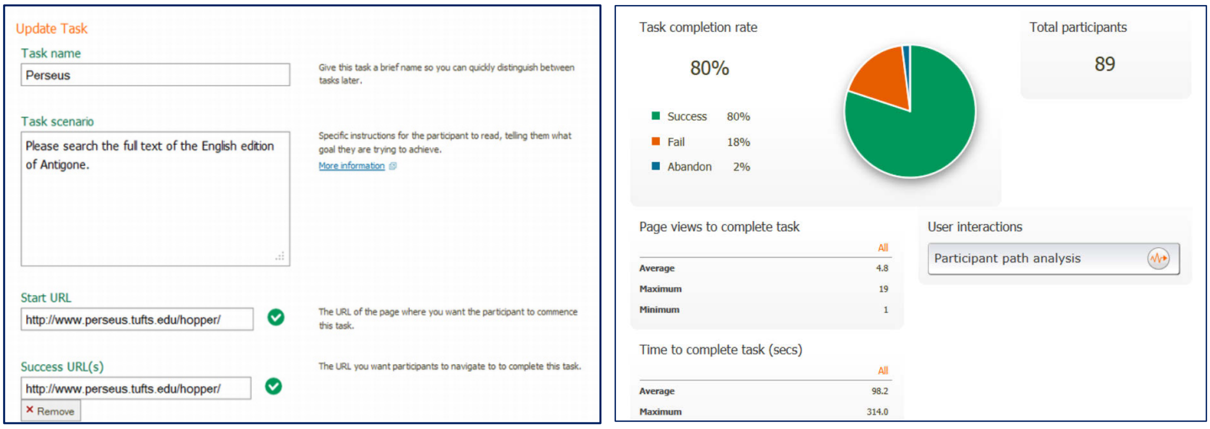
Several products are now on the market, sometimes even in forms that include an all-inclusive packet with participant recruitment and ready-to-go-data-interpretations. Librarians hesitate to make much use of these products, because they need (or want) to have low-cost tools. For example, Symonds (2011) decided on a self-made tool and used *SurveyMonkey* to build an asynchronous remote setting, in which participants received a survey link plus the request to open a website and keep that window open for the whole test. Participants had to go back and forth between the task description and test execution window and "users had to type how they had searched for information" (Symonds, 2011, p. 443). Symonds' article certainly admits the flaws of that approach, but concludes rather positively that this is a low-cost approach that might not be perfect, but at least collects valuable data.

Again, if the aim of a usability study is to detect problem areas, such an approach is fine, but for valid and reliable data collection in online environments, this approach has too many flaws. The participants have to switch windows between the task description and the websites themselves, and they have to describe their search behavior, which results in claimed behavior and also makes the questionable assumption that users were able to describe their search behavior in an accurate way. And finally, it reduces the potential participants to a set of very motivated users who are willing to take on this double effort. Participants with too much motivation can lead to a sampling bias and therefore to invalid data for hypotheses testing.

The following description draws on existing tools and how they allow asynchronous remote usability tests. During an asynchronous remote usability test, participants access a digital service and engage in small tasks, as in a traditional usability test. They read the task in written form next to, or above or below, the digital service or websites (see for an example appendix 3). When participants have navigated to the page that contains the information for the task, they can select "task complete" or they may click on "abandon task". Participants are also able to click on "task complete" without the relevant information. Their answer is then marked in the data as a "task failure". Some products allow participants to type the answer in an open form or choose between several options. Researchers can follow the participants' paths through click tracking. In addition to the retrieval tasks, researchers can also ask questions.

In contrast to usability tests that use small numbers of participants in synchronous tests and result in qualitative data, asynchronous remote usability tests use larger numbers of participants asynchronously and produce mainly quantitative data. A researcher learns how many participants solved a task and how many abandoned a task. They learn how much time participants needed for a specific task and at which point most participants gave up. These tests also allow an in-test-comparison of several digital services.

As with surveys, the technical set-up of asynchronous remote usability tests is extremely easy and quickly done, and the results are visible in real-time and downloadable in raw form or already processed for presentations (see figure 1).



Update Task

Task name: Perseus

Give this task a brief name so you can quickly distinguish between tasks later.

Task scenario: Please search the full text of the English edition of Antigone.

Specific instructions for the participant to read, telling them what goal they are trying to achieve.

More information

Start URL: http://www.perseus.tufts.edu/hopper/ ✓

The URL of the page where you want the participant to commence this task.

Success URL(s): http://www.perseus.tufts.edu/hopper/ ✓

The URL you want participants to navigate to to complete this task.

Task completion rate

80%

🟩 Success	80%
🟥 Fail	18%
🟦 Abandon	2%

Total participants

89

Page views to complete task

	All
Average	4.8
Maximum	19
Minimum	1

User interactions

Time to complete task (secs)

	All
Average	98.2
Maximum	314.0

Figure 1. Example of a task design and real-time result of an asynchronous remote usability test using *Loop11*.

The market for professional software solutions for asynchronous (and synchronous) remote usability tests grows each month. This research used the product *Loop11*, because it offered the possibility to test a type of asynchronous remote usability test that collects only quantitative data during the tasks. In addition, *Loop11* does not require an additional software installation on the participant's side. The other big players on the market at the moment are *Userzoom*, *Usertesting*, *EasyUsability*, *Webnographer*, *Mikogo* and *Usabilla*.

2.1.3 Limits of online studies

Online studies have many advantages and there is a clear need for them. However, they also have some limits, which could be a danger to the validity of the studies if not taken seriously. These are well documented in the literature and exist in addition to the question of the influence of distraction in a natural environment.

A representative study allows researchers to make statements about a specific population. But what is the population of a digital library? Some studies have tried to define the population of a digital library by developing personas (Akselbo et al., 2006) or to define the population as the "primäre Nutzergruppe" (primary user group), meaning the users who are registered for a local library (BIX Handbuch-WB, 2011, online). In contrast to physical libraries, digital libraries lack data about their population. The simple statement that 3,000 participants took part in a study does not contain any indication about whether the study is representative. The 3,000 participants could well be only very active or very satisfied users. Without knowing the exact digital library population, no sampling can be truly representative. This is probably the biggest flaw of online studies: a representative investigation is almost impossible.

This flaw results in convenience samples. Researchers draw on the users who are available and willing. The large majority of online studies, in particular in library and information science, use convenience samples. Researchers ask users on a website or via mailing list if they want to participate in a study. Through this approach researchers do not actively choose a sample, but let users decide based on willingness or interest in the study. This self-selection leads to an imbalance with highly motivated or especially interested participants. For example, some users might have a problem with a digital library interface and be more motivated to comment on it than users who have no problem at all. Synchronous remote tests reinforce that bias by choosing participants on availability of certain technology: headsets, for example.

Several approaches reduce the risk of a sampling bias. Instead of asking every website user, a pop-up-option after each nth user might provide a better sample. The social sciences also use panels. A panel involves a group of users who agreed to take part in studies on a regular basis. Panel members are actively recruited in order to reach a maximum level of representativeness of internet users. They allow researchers to define the characteristics of users and to use an accurate sample.

Critics say that panels have too many incentive-hunters, whose aim is to make money rather than to participate in a serious way. This is also true for recruitment on websites or mailing lists. Alternative recruitment techniques are slowly emerging like sampling using social networks (Baltar, 2012), but more research needs to be done to demonstrate the validity of these techniques on a broader basis.

Legal and ethical aspects establish certain important boundaries of online studies. The internet is by and large an open space and an El Dorado for researchers in the sense that it was never before so easy to examine so many different places without having to travel around the world. However, researchers are legally prohibited and should ethically refrain from examining everything that happens on the internet. This is also true for online studies in digital libraries. Digital libraries are public spaces, but users within digital libraries might (rightfully) perceive their behavior as private.

> “[J]ust because people's expressions on the internet are *public* in the sense that they can be viewed by anyone does not mean that people are behaving as though their audience consists of billions of people across all space and all time. How we act in a park with our children is different from how we act in a pub with our friends; just because these are both public places does not mean that there is a uniform context. When we look to understand people's practices online, we must understand the context within which the individuals think they are operating. This imagined context provides one mechanism for bounding our research.” (Hine et al., 2009, p. 31)

Another risk to data validity is the high number of drop-outs during an online user study. Few participants leave a laboratory after half of the test, but in an online user study participants need only one click to leave. Stieger & Göritz (2006) reported that 10.5% of participants left the chat during a synchronous interview and Tullis et al. (2002) reported a 20% drop-out rate during an asynchronous remote usability test. The experiment described in this dissertation had a drop-out rate of approximately 10%.

2.2 Test settings and influences on behavior

2.2.1 Laboratories and laboratory effects

A laboratory is an artificial environment created by a researcher. It can have very different meanings in different areas: usability researchers define laboratories as the place where they run usability tests. These laboratories are not necessarily clinical or uncomfortable environments; for example, researchers often serve coffee or juice to make the test experience agreeable. In information retrieval, laboratories might be more formal, but also have little resemblance to the stereotypical image called upon the word “laboratory”. Laboratory studies in interactive information retrieval aim at observing a particular information-seeking behavior while reducing disturbing variables. Participants are given retrieval tasks and researchers observe the participants’ behavior. For psychological experiments participants are divided into groups: one group receives treatment and the other serves as control. A particular treatment is intended to allow researchers to define the effect of a single variable.

All three approaches have one characteristic in common: in a laboratory, researchers can control a situation. They can limit confounding variables and can plan the succession of specific tasks. The key benefits of laboratories are their ability to control the causality. If researchers keep one variable constant, they can argue that the manipulated variable is the cause for the change in behavior.

Laboratories have long been seen as reasonable proxies for user studies, but in the last decades, the critics of laboratories increased. Researchers refer to laboratory effects, which affect performance. In a laboratory, participants find themselves in an altered state of stress and do not perform the tasks “in the same manner during the test as they would in a familiar work environment. In addition, the presence of an observer or the feeling of being tested may create unnecessary anxiety or pressure to perform” (Andrzejczak & Liu, 2010, p. 1258). Martin Orne in 1962 characterized laboratory situations as having a “demand character”, because the participants are aware of being controlled. In consequence they try to be cooperative and helpful and do not behave in a realistic way, but follow the expectations and wishes of the researcher instead. This effect is also called the “subject expectancy effect”.

The Hawthorne effect (coined by Roethlisberger et al., 1975) presses this behavior even further. Roethlisberger and his colleagues discovered in an experiment on workers’ productivity at the

Hawthorne Works company that participants performed better when they knew they were observed than they did when they thought they were not being observed.

Critics reproach laboratories for not representing behavior as it exists in real life and demand that phenomena should be examined in their "natural habitat" including any interactions that occur in that environment (Frey, 1987). Laboratories make behavior visible, but they also reduce it to an artificial state. The complexity of human behavior is reduced to a small number of variables that exclude external impulses.

2.2.2 Natural environments and influences due to distraction

Laboratory tests can result in better performances than users might show in a real-life situation. However, it is not clear if the reverse is true; that is, that in a natural environment test setting participants would perform with a lower level of effort than they would in a laboratory. Testing in the natural environment is still a test and not a real-life situation.

The natural environment is the every-day information-use-environment of digital library users. Some researchers call it the real life environment (for example Bowman et al., 2010). It is the setting in which users browse digital libraries, hang out, discover or start searches based on their own interest. The natural environment is simply the place where the users are—independent of whether this is at home, at work, with friends, at the office, in transit or somewhere else.

Today's world is noisy and so is the user's natural environment. As long as researchers do not explicitly collect contextual information about the natural environment, in which users interact during the study, behavior will be difficult to explain. Researchers can control variables in a laboratory, but they are unable to control the natural environment. While in a laboratory, the manipulated variable can be reasonably made responsible for a change in behavior, the natural environment lacks this easy causality. The effect of the independent variable can be confounded with other uncontrolled factors. This means that the advantage of testing in a natural environment includes the risk that disturbing events like phone calls occur during the test and then influence the results. Brewer (2000, p. 14) reported from a study in a natural environment that knowing about events like disturbances was an essential component of data collection: "The researchers were not only helpless to prevent such events but would not have been aware of them if they did take place".

The danger of data collection in a natural environment is not that events might occur, but that researchers know nothing about them.

It is a reality that distraction has become part of a users' life. Head & Eisenberg (2011) interviewed university students wanting to assess how students manage technology while in a library. They reported that in the hour before the interview "81% of the students in our sample had checked for new messages (e.g., email, *Facebook*, IMs, texts)" and discovered that the "most frequent combination (40%) of devices being used was a cell phone (including smart phones) with a personally owned laptop computer while they were in the library" (Head & Eisenberg, 2011, p. 3). Contacts like phone calls or SMS are undeniably part of the natural environment.

Televisions are another distracting element in the natural environment. In some home environments the device is always on. A study by Brasel & Gips (2011, p. 530) examined multitasking across television and internet content and discovered that "participants switched between media at an extremely high rate, averaging 120 switches in 27.5 minutes". This study also offered evidence that participants significantly underestimated their own distraction by 88% in a natural environment. González & Mark (2004) discovered that people switched tasks on the order of every four to eleven minutes.

People underestimate their distraction, because they believe that they are able to multitask. Multitasking describes a situation in which an individual handles more than one task at the same time, like watching TV and doing homework. Multitasking refers to a situation "where a person has to complete multiple tasks, but cannot execute them sequentially (due to time limitations) or simultaneously (due to physical or cognitive limitations)" (Law et al., 2006, p. 28). Young people believe especially that they are capable of multitasking (Bowman et al., 2010). Some researchers even call these young people *Homo Zappiens*[2] and state that we "see children today doing their homework, watching YouTube, instant messaging (IM), Twittering, using FB, surfing websites, and so forth in a way that seems as if they are doing all of this simultaneously" (Kirschner & Karpinski, 2010, p. 1237–1238). But multitasking does not signify that people really can do several things simultaneously. It only means that these tasks are interleaved with one another "each being

[2] The term was originally coined by Veen, W. & Vrakking B. (2006). *Homo Zappiens: Growing up in a digital age*. London: Network Continuum Education.

suspended and then resumed after appropriate intervals" (Burgess, 2000 cited from Law et al., 2006, p. 28).

People multitask when they use digital libraries in a natural environment. Multitasking means that they switch between the digital library and at least one secondary task with each single switch being a small distraction. In that sense the natural environment is an environment of continuous partial attention. As long as the tasks are similar in character, the influence of these switches might be only slightly harmful. A contact, however, is a substantial disturbance during digital library usage, because it requires a mental shift from one action to a completely different kind of action (Brasel & Gips, 2011).

The influence of distraction on behavior is an ongoing research area in many psychological studies. This dissertation does not attempt to present an in-depth analysis of this discussion, but narrows it to a few studies that are relevant for the research design. Researchers agree that distraction has an influence on the behavior, but have different opinions on the concrete form of that influence. Fried (2008) found that divided attention leads to lower task performance and Adamczyk & Bailey (2004) supported this by stating that interruptions can affect task performance. They also showed in their study that interruptions had an impact on frustration and annoyance. People who were distracted were likely to give more negative ratings.

Law et al. (2006) elaborated on the question of performance and found that participants gave a higher priority to more engaging tasks and that these engaging tasks consequently had a negative effect on the performance of the primary task. Law adds that participants showed that behavior despite instructions to the contrary. This means for asynchronous remote usability test design that the instructions to close open programs or not to talk to someone may well be ignored.

The time to complete a test is a factor that is used in psychology to measure the level of distraction. Mark et al. (2008) used time scores as a measurement of distraction. They discovered that undisturbed participants took more time to complete tasks than while being disturbed. If participants were disturbed, however, they showed a higher level of frustration, time pressure and more stress. Mark et al. offer a possible interpretation for their result: "When people are constantly interrupted, they develop a mode of working faster (and writing less) to compensate for the time they know they will lose by being interrupted" (Mark et al., p. 110). The test condition for this experiment was writing emails in an office situation. Czerwinski et al. (2000) studied the effect of distraction on completion time scores. They observed two groups in a laboratory: one group was disturbed twice by an incoming chat. The participants had to search in a book list with two levels of difficulty: find specific book titles and find books on specific topics. They collected data on completion time, and removed the chatting time in the chat group.

Czerwinski et al. (2000, p. 361) made four important discoveries:

(1) distracted participants take more time to complete a task;

(2) this difference in task time between the two settings is statistically significant;

(3) the difference in time is not entirely related to changes between keyboard and mouse, but comes from the influence of distraction on memory;

(4) distractions "reliably harm faster, stimulus-driven search tasks more than effortful, cognitively taxing search tasks".

Similar research was done by Bowman et al. and Kirschner & Karpinksi. Bowman et al. (2010) discovered that there was a significant difference between participants who were interrupted by instant messaging during task completion and undisturbed participants. However, this difference was only on the performance score of completion time. Results showed no evidence for a difference in successful task completions. This result was supported by Kirschner & Karpinksi (2010), who stated that participants showed a similar performance but needed significantly more time to complete the tasks.

2.3 The Elaboration Likelihood Model

It is important to collect judgments like ratings in user studies. At least as important is the knowledge on which basis participants made these judgments. In 1986, Petty & Cacioppo developed a model for decision-making processes which they called the Elaboration Likelihood Model. This model argues that distraction leads to an attitude change in the decision-making process.

The Elaboration Likelihood Model is a well-established model in the social sciences and has lately regained new interest in the areas of persuasive technology and in library and information science (Fogg et al., 2003; Hilligoss & Young Rieh, 2008 or Lim & Simon, 2011). The model provides a framework “for organizing, categorizing, and understanding the basic processes underlying the effectiveness of persuasive communications” (Petty & Cacioppo, 1986, p. 125). Elaboration means “the extent to which a person thinks about the issue-relevant arguments contained in a message” (Petty & Cacioppo, p. 128). The model distinguishes two distinct routes in persuasion, the central and the peripheral route:

> “The first route, which we have called the ‘central route’, occurs when motivation and ability to scrutinize issue-relevant arguments are relatively high. The second, or ‘peripheral route’, occurs when motivation and/or ability are relatively low and attitudes are determined by positive or negative cues in the persuasion context which either become directly associated with the message position or permit a simple inference as to the validity of the message.” (Petty & Cacioppo, 1986, pp. 131–132)

People use the central route in decision-making processes when their decision is a result of issue-relevant argumentation. Petty & Cacioppo call these elements of persuasion “arguments”. An attitude change occurs if there are elements in the contexts that would prevent people from scrutinizing arguments. In this case, people will take the peripheral route and base their decisions primarily on so called “cues”.

The model predicts that if people are motivated and able to scrutinize information, they will take the central route. If in the process of making a decision external stimuli are present, people will take the second peripheral route (more general information about the model can be found in Petty & Cacioppo (1986) and in Frey & Stahlberg (1993)). The model is not without criticism (see for an overview Perloff, 2003), because the cues on the peripheral route are perhaps too broadly defined. For the purpose of this dissertation, it demonstrates a possible influence of distraction in the natural environment on judgments.

Translated to the digital library world this means that under ideal conditions users base their judgments on arguments like search functionality or relevance of search results. Ideal conditions mean that users are motivated to use a digital library and able to scrutinize information. There is arguably no such thing as an ideal condition, in which information-seeking takes place in a digital library, but by excluding disturbing factors, laboratories aim at coming close to such ideal conditions. Of course researchers know that in natural environments, other things like satisficing, time constraints or personal design preferences drive users’ decisions.

The model supposes, however, that if users are in a natural environment setting, they are more likely to take the peripheral route and base their judgments on cues like design or the professional appearance of a digital library. This is because of the strong influence distraction has on decision-making. Distraction affects “a person's ability to process a message in a relatively objective manner. Specifically, distraction disrupts the thoughts that would normally be elicited by a message” (Petty & Cacioppo, 1986, p. 141). Distraction makes people use other information and as described in section 2.2.2, distraction is an elemental part of the natural environment.

The Elaboration Likelihood Model groups laboratory participants into an ideal condition that does not exist in reality. Even in a controlled environment like a laboratory, the likelihood that participants base their decisions only on arguments is rather small, because the conditions in a laboratory are never as ideal as the model supposes. Still, the model is a useful mechanism to examine the influence of distraction. Even if the laboratory setting does not offer ideal conditions, there should be a difference in decision-making between the laboratory and the natural environment, where distraction plays a much larger role in the latter. The aim of applying the Elaboration Likelihood Model is to see whether the natural environment setting is a factor that leads to a specific attitude. It is argued that distraction leads to the use of cues and less use of arguments. The Elaboration Likelihood Model is a useful step towards discovering if the judgments made in both settings appear to be the same, but actually are based on different things.

2.4 Earlier investigations on laboratory and remote settings

Online studies are one way of dealing with the limitations of laboratories. Because of the need to test at distance and to produce valid data with new methodological approaches, many researchers have conducted comparative studies. At least 30 studies have been performed to examine whether one can replace laboratories with remote settings (Bruun et al., 2009). The results of these studies vary as do the research designs.

This section presents selected studies to illustrate the way comparisons were done, and discusses how much this dissertation can build on earlier studies. The experiment described in this dissertation compares a laboratory setting with a natural environment setting using a remote approach. It does not compare remote and laboratory.

Some studies compared a laboratory setting and a single “remote” setting (like Kelly & Gyllstrom, 2011 or Dixon, 2009). Others compared the laboratory setting with several different forms of remote settings (like Bruun et al., 2009 or Andreasen et al., 2007). The test objects were software products (*Thunderbird*, an MP3 player, or an *Eclipse* plugin) or company websites; only Kelly & Gyllstrom (2011) used an information retrieval test collection.

The tasks were in general specific, but no study used tasks with clear endpoints that a system could detect automatically (like success URLs). A clear endpoint is a necessity for automatic measurement of task completion, which in consequence no study was able to offer. Instead, successful task completions in the remote settings were measured by post-test questionnaires (for example Thompson et al., 2004; West & Lehman, 2006; Mankoff et al., 2005) or self-reporting mechanisms during the test (Bruun et al., 2009; Andreasen et al., 2007). Bruun offered participants a hint that allowed them to check the correctness of their solution. In the laboratories, most settings had an in-room moderator who noted the result and participants were usually invited to think aloud.

The leading question behind these studies was the similarity of data in different settings. Batra & Bishu (2007) as well as Selvaraj (2004) found that remote usability testing was not different from traditional usability testing, while Andreasen et al. (2007), Brush et al. (2004), Petrie et al. (2006) and Bruun et al. (2009) discovered that fewer usability issues were identified in the remote setting. Andreasen et al. also found that there was no difference in task completion between the settings, whereas Tullis et al. (2002) discovered that there was a difference in task completion. Thompson et al. (2004) and Andreasen et al. remarked on a significant difference on the time spent on tasks between laboratory and remote, while Tullis et al. discovered that there was no difference on the time on task completion.

The strong differences between the studies seem strange, but they are actually unsurprising. All studies appeared to measure the same thing, when in fact they did not. In addition, while many of these studies have gone through peer-reviewing processes like the *CHI* proceedings*,* some research designs fail to persuade readers about their data validity. For example, it is no surprise that participants in the remote setting needed more time for task completion, because they had to type the answers and the participants in the laboratory only had to speak them aloud to the moderator sitting next to them.

Participants in the remote setting had to write about their behavior, which means that the data were only claimed behavior. Hence, researchers measured claimed behavior in the remote setting against actual behavior in the laboratory. Survey data on behavior reported the results of task completion, but the surveys missed information during the task solving process. Bruun et al. (2009, p. 1623) stated succinctly: “As we have no data on the task-solving process in the remote conditions, we cannot explain this variation [ed. in the standard deviations between the settings]” and so did Tullis et al. (2002, p. 2): "The information that can be collected using this technique [ed. survey] is limited. Since the two browser windows are basically independent of each other, it is not possible to detect what pages the user visits in the main browser, or any interactions with those pages. Our information is limited to what the users report to us in the small task window, plus the elapsed time.“

The number of participants varied between the studies, but was in general very low for the parametric tests they used to compare groups. Most of these studies had samples that were smaller than 20 participants, which resulted in an increase of the likelihood of type 1 and type 2 errors. For example, Thompson et al. (2004) had only 5 participants in each setting, Brush et al. (2004) had 8 in the laboratory and 12 remote and West & Lehman (2006) had 17 in the laboratory and 13 remote. A few bigger studies exist like Polkehn et al. (2010)—which had 67 participants in the laboratory and 544 participants in the remote setting—and Kelly & Gyllstrom (2011) which had 30 in the laboratory and 39 remote.

The recruitment strategies varied between the studies, but also within a study. Kelly & Gyllstrom (2011) decided on a passive form of recruiting and used an email list. They allowed participants to choose their preferred setting. Tullis et al. (2002) draw a random sample from the telephone book of the company for the laboratory setting and sent an email to randomly selected employees for the remote setting.

The estimated test time and the incentives were also different between the settings. Participants had to spend 90 minutes in the laboratory (Tullis et al., 2002 and Kelly & Gyllstrom, 2011) but only 60 minutes (Kelly & Gyllstrom) and 45 minutes (Tullis et al.) in the remote setting. Kelly & Gyllstrom paid their participants in the laboratory the double prize with the explanation that the “differences in compensation were justified by the differences in effort to participate and time.” (Kelly & Gyllstrom, 2011, p. 1534)

However, the crucial problem in these studies is the diverse understanding of "remote". Batra & Bishu (2007), Dixon (2009) or Andreasen et al. (2007) defined their synchronous remote setting as a "simulated remote environment" (Andreasen et al., p. 1408), which was basically the room next door. Bruun et al. (2009), Brush et al. (2004) and McFadden et al. (2002) allowed their participants to be in their natural environment and communicated by screen sharing, but told participants that they had to install the necessary software before the test.

Kelly & Gyllstrom (2011) and Tullis et al. (2002) defined their asynchronous remote setting in the sense of participants' natural environment. But they tried to exclude distraction as much as possible: Tullis et al. offered a "pause button" so that participants could "'stop the clock' on a task if they were interrupted or just wanted to take a break" (Tullis et al., 2002, p. 6). They also deleted all data where the individual task completion times were under 5 seconds or over a 1,000 seconds, because they interpreted time data under 5 seconds as an indication that "the user did not seriously attempt the task" and over a 1,000 seconds as an indication that "the user was probably interrupted" (Tullis et al., 2002, p. 3). Kelly & Gyllstrom told their participants that "they should complete the study in one uninterrupted session, close all other applications on their computers and not multi-task". Participants were not allowed "to answer their cell phones and/or read/send text messages" (Kelly & Gyllstrom, 2011, p. 1534). Instead of systematically deleting participants, Kelly & Gyllstrom discouraged participants from distraction. Participants were told that their keystrokes would be tracked and that "the system would automatically log them off after a 10 minute period of inactivity and they would not be able to resume the study later" (Kelly & Gyllstrom, 2011, p. 1534). This notion of "remote" resembles virtual laboratories, in which distraction is considered a confounding variable and is consequently eliminated.

Andreasen et al. (2007) and Bruun et al. (2009) faced the problem of distraction when they realized that without information on distraction, their data was difficult to interpret: "we do not know if the test subjects had any breaks during the test sessions, and therefore we do not know the exact time spent on the test" (Andreasen et al., 2007, p. 1410) or Bruun et al. stated that "the consequence is that we have missed information about their task solving process. It also means that the task completion times have to be read with great caution" (Bruun, p. 1625).

The difference between the approaches described in the earlier studies and the approach in this dissertation is not the factor “remote”, but the difference between natural and artificial (=laboratory) use environment. The experiment in this dissertation describes a remote test that has specifically been adapted to embrace the users’ natural environment. The natural environment can take on many forms of information spaces, but it is never simply “a transplanted replication of laboratories” (Brewer, 2000, p. 14).

2.5 Summary

Online studies have a dual relationship with the internet: the internet is both methodological tool as well as object. The characteristic trait of online studies is a spatial distance between researcher and participant. This trait is labeled remote. In a synchronous remote setting, researchers and participants are separated in space, but they have a real-time connection using text, voice or video. Asynchronous tests appeared more recently and add a temporal dimension. Researchers and participants are now separated in time as well as space. Participants have no direct contact with the researcher and can access the test at a place and time of their choice.

Usability designers were in a quandary: they valued their traditional approaches while at the same time seeing the need to test in online environments. For the last ten years, web surveys were the most popular method in online studies followed by log file analyses. Since 2008, remote usability tests have gained in importance. During an asynchronous remote usability test, participants access a digital service and engage in small tasks, as in a traditional usability test. In contrast to usability tests that use small numbers of participants in synchronous tests, asynchronous remote usability tests use larger numbers of participants asynchronously and produce mainly quantitative data. A researcher learns how many participants solved a task and how many abandoned a task. They learn how much time participants needed for a specific task and at which point most participants gave up. These tests also allow an in-test-comparison of several digital services.

However, online methods also have some limits, which could be a risk to the validity of the studies if not taken seriously. For example, a representative investigation is almost impossible and frequently results in convenience samples. Another risk to data validity is the high number of drop-outs during an online user study.

This work examines two research settings: a laboratory and a natural environment setting. In a laboratory, researchers can control a situation. They can limit confounding variables and can plan the succession of specific tasks. The key benefits of laboratories are their ability to control the causality. If researchers keep one variable constant, they can argue that the manipulated variable is the cause for the change in behavior. Laboratories make behavior visible, but they also reduce it to an artificial state. The complexity of human behavior is reduced to a small number of variables that exclude external impulses. Researchers criticize that laboratory effects, like the "subject expectancy effect", affect performance. The natural environment, on the other hand, is the every-day information-use-environment of digital library users. It is simply the place where the users are—independent of whether this is at home, at work, with friends, at the office, in transit or somewhere else. Researchers can control variables in a laboratory, but they are unable to control the natural environment. As long as researchers do not explicitly collect contextual information about the natural environment, in which users interact during the study, behavior will be difficult to explain.

Today's world is noisy and so is the user's natural environment. People multitask when they use digital libraries in a natural environment. In that sense the natural environment is an environment of continuous partial attention. The danger of data collection in a natural environment is not that events might occur, but that researchers know nothing about them. Researchers from psychology discovered in various studies that distraction has an influence on the behavior, but this influence can take on different forms.

In online user studies in digital libraries, it is important to collect judgments like ratings. At least as important is the knowledge on which basis participants made these judgments. In 1986, Petty & Cacioppo developed a model for decision-making processes which they called the Elaboration Likelihood Model. This model argues that distraction leads to an attitude change in the decision-making process. People will use the "central route" in decision-making processes when their decision is a result of issue-relevant argumentation. Petty & Cacioppo call these elements of persuasion "arguments". An attitude change occurs if there are elements in the contexts that prevent argument scrutiny. In this case, people will take the "peripheral route" and base their decisions primarily on so called "cues". The aim of applying the Elaboration Likelihood Model is to see whether the natural environment setting is a factor that leads to a specific attitude. It is argued that distraction leads to the use of cues and less use of arguments.

This work is not the first one that is interested in synchronous and asynchronous remote usability tests. At least 30 studies have been performed to examine whether one can replace laboratories with remote settings. The crucial problem in these studies is the diverse understanding of “remote” which ranges from the room next door to an online environment that resembles a virtual laboratory. Some researchers defined their asynchronous remote setting in the sense of participants’ natural environment, but tried to exclude distraction by adding a pause button or by telling participants that they must close all other applications and must refrain from talking. The results of these studies vary as do the research designs. Selected studies were presented to illustrate the way comparisons were done, and to discuss how much this dissertation work can build on earlier studies.

3 Pilot test design and results

This chapter describes a pilot test, which was developed and run in November 2010.[3] The aim of conducting a large pilot test with 31 participants was to explore the nature of the data that the program *Loop11* produces, to test the tasks and questions and to test the experimental setting. Detailed information on the tasks and the choice of digital libraries will be presented in chapter four. Screenshots of the complete pilot test can be found in the appendix 4.

The sample for this study consisted of library and information science masters' students at the *Berlin School of Library and Information Science.* All participants already held a bachelor's degree in library and information science and were expected to know how to search in digital libraries. 23 females and 8 males with an average age of 26.6 years participated in the test. The students were randomly assigned to two different test settings: one sample in a laboratory, the other in a place the students chose.

An email invitation was send to two working groups within a course using the e-learning platform *Moodle*. The email told the participants that the aim of the study was to test the usability of several digital libraries. They were explicitly told not to tell the others about the experiment. This ensured that neither group knew that there was both a laboratory and a natural environment setting. The laboratory group got information about when to come to the laboratory, where they would get instructions; the natural environment group received a direct link to the test with instructions. The instructions asked participants to do the test in their current environment: they were explicitly informed that they need not close any applications or to refrain from talking.

Both initial groups were similar in size, but not all students ultimately participated in the test. In the end, 13 participants completed the test in the laboratory and 18 participants completed the test remotely in their natural environment at a time of their choice. There was no significant relation between the hour when participants in the natural environment group completed the test and the completion time that they spent on the test. Both groups underwent an asynchronous remote usability test using the software *Loop11*. Each test was accessible by means of the internet; the

[3] This chapter is an adapted version of Greifeneder, E. (2011b). The impact of distraction in natural environments on user experience research. In: Proceedings of the TPDL'11 Lecture Notes on Computer Science (pp.308-315). Berlin, Heidelberg: Springer.

laboratory consisted of a computer pool. A moderator ensured that there was no external distraction and sat at the front desk. The situation resembled an exam.

The participants were asked to do similar tasks in each of five different digital libraries. These were *DigiZeitschriften*, *Social Science Open Access Repository* (*SSOAR*), *Perseus Digital Library (Perseus), Open Repository Kassel* (*ORKA*) and *Valley of the Shadow.* Except *ORKA* and an additional control website, these digital libraries were identical with the final test. In each digital library, participants had to search for a specific document. For example, one of the tasks was to search for a talk entitled "Demokratie durch Krieg". The task description was shown on top of the screen; the system being tested was shown on the rest of the screen. Participants did not have to switch between windows to see the questions and the digital libraries (see appendix 4 for illustrations). There was no need to install any additional software. All digital libraries were fully functional within the test window. Participants' confidence with the kind of digital libraries and the types of tasks was high, because the tasks resembled their preparations for essays or class papers.

After each task, participants rated the perceived difficulty of the task by judging if this task was "very easy", "easy", "neither easy nor difficult", "difficult" or "very difficult". At the end of all tasks, participants completed a questionnaire about themselves and their environment during the test, especially about their level of distraction. A 10-point likert scale was used.

The five digital libraries included three German ones—*DigiZeitschriften*, *SSOAR* and *ORKA*—as well as two English ones—*Perseus* and *Valley of the Shadow*. The choice for these digital libraries was based on the research design necessity that participants should not require high domain specific knowledge for task completion. *ORKA* and *Valley of the Shadow* were chosen, because the tasks required a higher cognitive load compared to the other tasks. The aim was to test if the assumption made by Czerwinski et al. (2000) was accurate that higher cognitively-loaded tasks result in no obvious difference between distracted and non-distracted participants.

The important information to gather in this pilot test was whether participants need more or less time in one of the two settings. The research's goal was not to analyze or to compare the usability of the five digital libraries.

In summary, there were several differences between the two settings in the amount of time participants needed. The mean *test duration* in the laboratory—i.e the time spent on the whole test—was 665.15 seconds; in the natural environment it was 1173.44 seconds, which is a clear indication of distraction in the latter setting.

An independent-samples t-test was conducted to compare the *test duration* in the laboratory and in the natural environment. There was a significant difference in scores for laboratory ($M = 665.15$, $SD = 153.73$) and natural environment participants ($M = 1173.44$, $SD = 845.01$; $t(18.54) = -2.50$, $p < .03$, two-tailed). There was no significant difference in scores for the number of *page views* needed to complete the test within the laboratory ($M = 28.85$, $SD = 8.30$) and natural environment ($M = 37.11$, $SD = 14.45$; $t(29) = 1.62$; $p > .10$, two-tailed). This indicates that although people in the natural environment needed more time to complete the test, they did not have more trouble searching the documents. The number of *page views* to complete the tasks is nearly the same in statistical terms. The reason for the longer *test duration* must lie elsewhere.

An additional factor between the two groups that has not yet come up in the literature is the large variability in *test duration* in these two settings. In the laboratory, there was a small variability between the five lowest scores (from 456 seconds to complete the whole test to 607 seconds) and the five highest scores (ranging from 682 seconds to 966 seconds). The natural environment group revealed a large variability: the lowest score (473 seconds for the whole test) was equivalent to the lowest one in the laboratory; but between the three highest scores in the natural environment lay more than 1,000 seconds (the three participants that needed the longest took 1,593 seconds, 2,731 seconds and 3,693 seconds). The laboratory shows an ideal user's behavior in which external factors were removed whereas time scores in the natural environment fluctuate heavily.

Based on earlier findings by Adamczyk & Bailey (2004) it was expected that participants in the laboratory would rate the digital libraries more positively compared to participants in their natural environment, who give more negative ratings. Table 1 indicates that this hypothesis might be true. The ratings in the laboratory tended to be more positive with more participants rating tasks as "easy". The two possible answers "very easy" and "easy" were combined under a joint measurement "easy" and the two answers "very difficult" and "difficult" were combined to a joint measurement "difficult". In the natural environment group, more participants rated tasks negatively, that means they regarded more tasks as "difficult" or as "neither difficult nor easy".

Options	**DigiZeitschriften**		**Perseus**		**SSOAR**		**ORKA**		**Valley of the Shadow**	
	LAB	NE	LAB	NE	LAB	NE	LAB	NE	LAB	NE
easy	100%	88.9%	69.2%	83.3%	100%	94.4%	69.2%	50.0%	15.4%	16.7%
difficult	–	11.1%	–	–	–	5.6%	23.1%	33.3%	53.8%	55.6%
neither difficult nor easy	–	–	30.8%	16.7%	–	–	7.7%	16.7%	30.8%	27.8%

Table 1. Participants rating the perceived difficulty of the tasks in the five digital libraries in the laboratory (LAB) and in the natural environment (NE).

The phenomenon described by Czerwinski et al. (2000) was mirrored in the results: for the most cognitively loaded task (here *Valley of the Shadow*), the difference between the two groups became visibly minor. This effect might be the result of a high engagement level during difficult tasks that reduces distraction. Because of the choice for categories instead of a Likert scale, statistical tests for comparing groups like the t-test were impossible. The results of the pilot test only indicate tendencies, but cannot give statistical evidence.

The pilot test also demonstrated that participants in the natural environment group were distracted during the test. 77.8% of the participants in the natural environment had programs open during the test; of these, only 33.3% said that they never looked at the open programs during the test. 38.9% of participants said that someone had talked to them during the test (also via SMS or phone). One question for both setting groups was whether they had been distracted by daydreams such as thinking about their shopping lists or class preparations: 74.2% admitted that they had been distracted by daydreams. About half of the whole sample said that they focused 90% to 100% of their attention on the test.

The experiment in this dissertation assumes that distraction is the reason for higher time rates in the natural environment group, but since the group in the natural environment was small (n = 18), no statistical tests to validate this hypothesis could be undertaken.

After the experiences with the pilot test, a few changes were made such as refraining from informing participants explicitly that they need not close any applications. The laboratory setting was changed from an exam character to a quiet reading room experience, because of the potential negative pressure on participants. A wider ranged and more complex system of judgments was developed. Participants stated that they found the 10-point-likert scale rather confusing and that they experienced the question about their social background (figure 69) as annoying. The digital library *DigiZeitschriften* was changed to be a training task, because it offered no permanent URLs which were necessary for a reconstruction of a participant's task completion. The digital library *ORKA* was excluded, because it went through a redesign and a server change just before the final experiment was scheduled. A control website that all participants knew was added to the test. Participants reported about technical problems during the test, so an additional question on these was added. Questions on participants' mood, on prior experience with the test objects and questions on language skills were added as well. Unclear question formulas were changed. Chapter 4 gives details on the test setting of the final experiment.

4 Procedures

4.1 Null hypothesis and alternative hypotheses

This experiment sought to examine whether the data gathered in a participant's natural environment differs from data from the same test when run in a laboratory, and if that difference can be explained by distraction. In what follows, participants' responses, judgments and behavioral manifestations are generally referred to as "data" where appropriate. Following the standard assumption in most asynchronous remote usability studies, the research design examines the following null hypothesis:

> There is no difference in the data gathered from the same test in a laboratory and a natural environment setting.

The design expects that the null hypothesis can be accepted in many or even most cases. The goal of testing broadly is to look for those particular instances where the null hypothesis has to be rejected without making prior assumptions about which instances these will be. The corresponding alternative hypotheses, which will be tested below, build on the Neyman-Pearson's test logic of statistical significance testing. In what follows they are referred to as hypothesis (in contrast to the central "research hypotheses" mentioned earlier in this work).

The experiment in this dissertation examines the following six hypotheses. It is expected that between a laboratory and a natural environment setting

(hypothesis 1) there is a difference in the time participants needed to complete a test;

(hypothesis 2) there is a difference in the variability in the time participants needed to complete a test;

(hypothesis 3) there is a difference in the participants' judgment of the digital libraries;

(hypothesis 4) there is a difference in the decision-making process for participants' judgments.

(hypothesis 5) there is a difference in the number of page views participants needed to complete the tasks;

(hypothesis 6) there is a difference in the number of successful task completions.

4.2 Research design

The research design differs from other studies that compared user behavior in a laboratory and in a remote setting. Earlier research (see chapter 2) tried to reproduce a remote setting that resembled a laboratory, but in a virtual environment with spatial and temporal distances. The remote test situation here intentionally embraces all factors that are part of the user's real natural environment.

The analysis of earlier studies also showed that some issues in the research design are of great importance in order to ensure the comparability between the settings. These are: an identical recruitment strategy without revealing the two different settings, an identical estimated time on the test, and identical incentives.

Participants were randomly assigned to two samples in similar size from the same population in which one sample, the experimental group, received treatment and the other group, the control group, received no treatment. The treatment in this study was the test setting. The aim of this design was to maximize the relative amount of variance due to the independent variable.

The chosen laboratory setting in this work consisted of a small computer pool inside the university library. Recruited participants for the laboratory were encouraged to come to the computer pool immediately, but were given the opportunity to show up later or even on another day. Participants in the laboratory were not allowed to talk or to distract themselves in any visible way. There was always at least one recruiter watching the participants in the laboratory. The situation resembled an extraordinarily quiet reading room. The laboratory was easily accessible and consisted of 15 computers with *Windows* as operating system. Participants in the laboratory used the default university browser *Firefox* to access the test. In the laboratory, the start page on the screen consisted of a white page with a short written text: “click here for the test”. Once participants clicked on the link behind the word “test”, the main page of the experiment opened. This construct aimed at an identical starting point for both the laboratory and the natural environment. Participants in the natural environment had to type the link into the browser. With the extra page “click here for the test”, both groups had the same chance to decide how much time they wanted to spend on the first instructional site of the real test, for which, of course, the time clock started to run.

The remote situation was different: no demands were placed on where and when the test should be accomplished. In general, remote participants were told that they had three days to take part in the test to encourage participation. Recruiters informed participants that they could do the test at a time and place of their choice and on a machine of their choice. Participants received a link on a paper sheet and were asked to visit that website. They were not given any other explicit restrictions: for example, they had no instructions to close any running programs and no ban on talking to others. The aim was to have a realistic online test environment that resembled the user's everyday digital library use environment.

The financial reward was the same in both settings: a five Euro voucher for *Amazon.de* which participants received by email address (there was no requirement to provide the email address but most did). The *Berlin School of Library and Information Science* supported the vouchers, which were chosen as an easy way to distribute incentive money to online participants.

Participants were told that the aim of the study was to improve the usability of digital libraries. This was necessary for the research design, because participants should not know this was an experiment about distraction. They did not know that their time was being measured or that the study was about distraction in a natural environment setting. Questions about distraction came at the end of the test with an assurance that the participant's answers had no influence on their reward. Personal data—in this case the email address for the rewards—was immediately separated from the data set and stored as an encrypted file on an external hard disk. For this test, the software *Loop11* was used. *SPSS* was used for data analysis and the data input in *SPSS* was completely anonymized.

Earlier research projects on remote test settings collected data on the variable *test duration*, that is, on the time to complete the whole test. This number was usually gathered for time on task performance measurements. Few studies assembled the time spent on the tasks (referred to as the variable *time on tasks)* and the time spent on the questions (referred to as the variable *time on questions)* separately. Most studies collected a single number for duration ("Time taken to complete the study" (Kelly & Gyllstrom, 2011) or "the time spent conducting the usability tests" (Bruun et al. 2009)). This study collected data on *test duration* as a central number and in addition the variables *time on tasks* and *time on questions*. Other studies additionally collected data on page views, successful task completion, perceived difficulty of a task, and ratings of websites, which the present study collected as well. A clickstream record of each participant was provided.

Time is frequently interpreted as an implicit measure of interest (Kellar et al., 2004 or Claypool et al., 2001) with longer completion time scores indicating an interest. But this interpretation fails if participants get distracted during the test. Longer completion time scores then may indicate the contrary: a potential lack of interest that might have left them open to distraction. Instead of collecting information about *test duration* to measure time-on-task-performance or to measure an interest, *test duration* was used as an indicator for the existence and the amount of distraction. This approach is based on the results of Czerwinski et al. (2000). Completion time is therefore a tool to demonstrate distraction.

The online test was pretested by a small group of participants before the pilot test. The final version described below was again pretested with students at the Humboldt-Universität zu Berlin, students at other German universities, and with the author's friends and colleagues.

4.3 Recruitment process

The recruitment took place in the lobby of the Jacob-und-Wilhelm-Grimm-Zentrum, the main university library building of the Humboldt-Universität zu Berlin. In total, 84 participants were successfully recruited. About 5000 students enter the library each day. Many of them spend time only in the lobby. The university library gave official permission for the experiment and the library security staff was notified about the recruitment.

The recruiters were student assistants from the *Berlin School of Library and Information Science* plus the author who is a full time faculty member at the same school. Several recruiters had experience with advertising work or telephone interviews. In total, eight recruiters were involved. The high number of recruiters and their individual approaches guaranteed a well-balanced group of participants. Each recruiter received detailed instructions in written form. A recruiter never worked for more than an hour; after that the active recruiter switched with a second recruiter sitting in the laboratory.

A random assignment was used. Recruiters strictly alternated between participants for the laboratory and for the natural environment. When they asked library users standing in a group, they would allow two at a time for the same setting. Three or more participants at once for the same test were never allowed, because of a potential sampling bias.

In the most frequent recruitment form for online tests, participants volunteer on their own to take part in a test by clicking on a link that was put on a website or sent through an email list. In this so-called passive recruitment, self-selection leads to highly motivated participants and unbalanced groups. This experiment decided on an active recruitment among people who visit the library lobby. The recruiters asked actively and sometimes even persuaded to participate in the test. Because recruiters made the choice whom they were asking, the problems of incentive-hunters or over-motivated participants could be limited. The aim was to get a wide-ranging and well-balanced group of participants.

Recruiters used different tactics to engage potential participants. They asked users at the library entrance, in front of the lockers or in the smoking area outside the lobby. Recruiters asked at the book return machines in order to get access to the “drive-through-user”, who enters a library simply to return or pick up materials and then leaves as soon as possible. The cafeteria inside the lobby was also a good recruiting place, because many users seemed to be open to participating after a cup of coffee. Some of these cafeteria users were not frequent library users, because they only use the cafeteria in the lobby, and do not use the library services at all. This kind of user is very hard to engage with passive recruitment on a library website.

Recruiters had two different invitation sheets to distribute to potential participants (see appendix 2). The one for laboratory participants showed a picture of the lobby with an arrow directing to the computer pool and informing about the opening hours of the laboratory. The second sheet for participants in the natural environment provided the test link along with information when to complete the test. Participants in the natural environment group could volunteer to receive an electronic reminder by email that same evening with the link to the test.

Participants did not know that there were two settings: they never had the choice between laboratory and natural environment. Recruiters had a strict ban on exchanging a laboratory sheet with a natural environment sheet or vice versa if, for example, a participant was interested, but had no time to come to a laboratory on this day. This approach differs from other studies in which participants knew that there were different settings and sometimes could even choose their favorite setting—a choice that leads to biased results.

Any university student could participate in the test; high school students or Humboldt faculty where not allowed to participate. There was no limitation to a specific university. Berlin has three big universities and a number of smaller ones. The city of Potsdam with its universities is also close. The intention was that the participating students belong not only to the Humboldt-Universität zu Berlin and the recruitment spot is known to provide study space to students from all of these universities. Berlin is an international city and the universities have a large population of international students, who were allowed to take part in the test, too.

The recruitment took place from May 3–9, 2011. A first recruitment cycle was scheduled for Monday, May 3 from 2PM to 7PM and May 4 from 9AM to 12AM. A second recruitment took place Wednesday, May 5 from 2PM to 6PM, Saturday from 11AM to 5PM and Monday from 9AM to 4PM.

4.4 Tasks and questions

Participants had to complete small research tasks in five digital libraries in the following order: *Perseus*, Social Science Open Access Repository (*SSOAR*), Digital Picture Archives of the Federal Archives (*Bundesarchiv*), *Valley of the Shadow* and *Amazon*. It could be argued whether or not *Amazon* counts as a digital library, but for ease of reading, this work refers to all tested websites and services as digital libraries when referring to them as a whole. A training task was offered in the digital library *DigiZeitschriften.* The choice of the digital libraries and the tasks, which will be described below, was due to the fact that the research design needed test objects and test tasks that did not necessarily require prior knowledge of the digital libraries or the topics of the tasks. Since the population consisted of students from various disciplines, all participants from the population needed to be able to complete the tasks. This excluded tasks or digital libraries that required highly domain specific knowledge. The number of German digital libraries that fulfill this requirement and in addition provide permanent URLs is limited and led to the four digital libraries mentioned above.

For the analysis, it was important to have at least one website that was very likely known by all participants in order to see whether previous knowledge of the site changed the behavior. This control website needed to be similar to the previous digital libraries in the sense that it allowed a similar search task. *Amazon* was chosen as this control website, because it allowed users to search for a book, had permanent URLs and the search results were relatively stable compared to, for example, *eBay* or *Google*. With 16.7 Million users in Germany, *Amazon* is the number one online shopping website in Germany (GfK, 2009) and as such qualified as a control website that all participants were likely to know.

The small retrieval tasks were similar to finding information for a class paper: participants had to search for a concrete article, for a specific picture, for a book, and for a page number. If participants found the requested information, they could click on “task complete”, or if not, on “abandon the task”. Participants could also click on “task complete” without having found the relevant information. Their answer was then marked in the data as a task “failure”.

As shown in chapter 2, the usefulness of participant’s self-reporting of task completions was limited. Therefore, the experiment decided on an automatic way of measuring task completions. This approach meant that tasks needed clear, measurable endpoints. All tasks required participants to search for a concrete piece of information within a document, because the software *Loop11* needed concrete success URLs to define which task was a successful search. But *Loop11* only provided information about the success URLs (within the task) and not if an answer to a question was correct (within a questionnaire). Therefore participants were asked to find the information, but did not need to provide it explicitly after the task. This approach had the advantage that participants could abandon a task without feeling guilty twice: first at abandoning the task and second at being unable to provide the answer on the next page. The idea of Bruun et al. (2009) to use hints was adopted: all tasks offered a small hint so that a self-administered check was possible. For example, participants had to find the first verse of Antigone and the hint stated that it started with “Ism”. The first verse of the tragedy starts with Antigone’s sister speaking and her name is Ismene.

Participants had to complete different kinds of retrieval tasks at different levels of difficulty and in different languages. The pilot test indicated that the difference in participants’ judgment of a digital library between the two settings becomes minor if the cognitive workload of the tasks is high enough (see also Czerwinski et al., 2000). Because of its design, the pilot test was unable to give statistical evidence for that. The final research design included two difficult tasks to test the effect of task difficulty on participants’ judgments. This effect becomes visible if hypothesis 3 (supposing that there is a difference in participants’ judgments of the sites) is true for the three easier tasks, but not for the two more difficult tasks.

The research field of multi-lingual information retrieval and access recognizes that language plays a crucial factor in retrieval behavior. The influence of language could not be tested extensively within this test and is limited to two tasks in English-language based digital libraries and to non-native German speaking participants using German and English digital libraries. Sometimes a task was difficult, because the requested document was difficult to find (*Bundesarchiv*) or sometimes because the digital library required cultural knowledge that most participants lacked, which made the task difficult (*Valley of the Shadow*).

Especially in information retrieval, task randomization plays an important role, because of the effect one task can have on participants' behavior in following tasks (see for example Byström & Järvelin, 1995; Borlund & Ingwersen, 1997 or Cole et al., 2011). Yet in this work, the order of the tasks was deliberately not shuffled. Randomization of tasks reduces systematic influences on the data by producing unsystematic varying sources of error. Since much of the previous research on task order was done using within-subjects-designs (researchers examine one participant on task 1 and task 2), it is unclear if these results on task order hold for the chosen between-subjects-design (researchers examine one task completed by participant 1 and participant 2). Keeping the task order as a constant factor between the settings allowed maximizing the relative amount of variance due to the independent variable. The non-randomization also examined whether a specific order of tasks leads to a particular behavior in the natural environment. The test explicitly started with easy tasks and moved forward to difficult tasks and tasks in English (which was a foreign language for most participants); potential effects of this task order in the natural environment setting were examined. The schema for the tasks and choice of digital libraries can be seen in table 2.

	digital library	**difficulty of task**	**language of digital library**
training task	*DigiZeitschriften*	easy task	German
1st task	*Perseus*	easy task	English
2nd task	*SSOAR*	easy task	German
3rd task	*Bundesarchiv*	difficult task	German
4th task	*Valley of the Shadow*	difficult task	English
5th task	*Amazon* (control website)	easy task	German

Table 2. Order of tasks, language and difficulty.

A copy of the complete online test can be found in the appendix 3. The test started with an introduction page, welcoming the student and explaining the aim of the test as well as responsibilities. The test invited participants to imagine that they have to complete an assignment for a class session and need to find several documents as well as a picture to finish the homework. This scenario was well known to participants, because finding documents for classes is part of their regular university life. Participants were also informed in both conditions that the test would last approximately 20 minutes. The second page showed an explanation on how to use *Loop11*. The third page asked participants about their current location. Participants could choose between laboratory and “somewhere else” and were asked to specify where. This question was required.

Participants started with an easy task in the digital library *DigiZeitschriften,* which is the German equivalent to *JSTOR*. Participants had to search for a document called “Zur Wergeldfrage” and find the publishing date. This task was designed to make sure participants knew how *Loop11* works and what the tasks look like. Participants were also informed that they do not need to provide the answers to the tasks.

The time spent on the tasks started to be measured at the beginning of the next section, devoted to the digital library called *Perseus*. The task in this English-language art and archaeology digital library and the task in the digital library *SSOAR* were both straightforward. Both digital libraries provided a search box and trustworthy search functionality. In *Perseus*, participants had to search for the English version of Antigone and find the first verse of the tragedy. In *SSOAR*, participants had to search a document by Michael Ramm entitled “How to Improve the Course Situation in Natural Sciences” and to find out how many pages this document has.

The digital libraries *Bundesarchiv* and *Valley of the Shadow* were chosen, because the tasks required a higher cognitive load compared to other tasks. The task in *Bundesarchiv* forced participants to find a particular detail in a photo. The question stated that participants should search for a picture of the federal party convention of the German CSU (“Christlich Soziale Union”) dating back to 1973. Participants received additional information on persons represented in the photo (Helmut Kohl, Henry Kissinger, Karl Carstens, Kurt Biedenkopf, Ludwig Erhard), although this information was not necessary to find the photo. Most participants started a long query including all available information instead of starting with a minimum of terms. This user behavior was interesting in itself, but lies outside of this project's scope and will not be examined here. Participants were asked to discover what object was placed in front of the podium, offering the hint that it started with “Steuer”. The task in the *Bundesarchiv* was triply difficult: first, finding a picture is usually more difficult than finding text (see for example Rui & Huang, 1999). Second, the task description provided more information than needed: including all the information actually led to a fuzzier search result than searching without all the terms. Third, the hint led participants into a wrong direction, because the word “Steuer” has two meanings in German: tax and steering wheel. Participants generally expected an object related to tax, because the picture was about politics. However, the picture showed a steering wheel (“Steuerrad”). The aim of misleading participants was to make the task more difficult. The aim of that task was also to test if a background in history caused a difference in retrieval behavior.

The task in *Valley of the Shadow* was less difficult, but the digital library in itself is culturally complex. Historical knowledge about the civil war is very helpful in using the navigation. The participants had to search for a letter from Toni Pastor to Annie Harris dating back from March 8, 1864 and to find the location where he wrote the letter. While this task sounds fairly easy, the difficulty with this digital library is that the first search box option appears on the site's third sub-page and that the search functionality is not user-friendly.

The *Amazon* website served as control in order to have at least one website that was known to all participants. The task was meant as an easy finish and asked participants to find a particular book on how to write a scholarly paper by Frank Grätze published in 2006 and to find the number of pages that book had.

After each task, five questions identified the level of perceived difficulty of the task, the relevance of the search results, the search tool performance, the design and the professional appearance of the digital library. The aim of these questions was to test, if the expected level of difficulty matched the perceived level of difficulty and to test in how far the natural environment leads participants to choose another route in their decision-making process on judgments. This examination was based on the Elaboration Likelihood Model (see chapter 2), which predicts that distraction leads to the use of “cues” and less use of “arguments” in persuasion.

All questions asked the participants to provide their answers on an 8-point Likert scale. An equal number on the Likert scale was chosen explicitly to avoid errors of central tendency and to force an indication of the direction of the participant’s choice. A wide range of scores was selected in order to increase the number of possible statistical techniques, which then can be used for data analysis (for an extensive study on that issue see DeVellis, 2003). A scale of 10 (1 = I agree and 10 = I disagree) is frequently used, but seemed too much differentiation on a statement about mood such as "I feel good".

At the end of the experiment, participants were asked to offer information about their distraction level including open programs, contacts or technical problems. These questions were meant to confirm and explain the indications of distraction based on completion time. In order to avoid social desirability responses, participants were told that honest answers were important for this research and that their answers had absolutely no consequences for their reward.

4.5 Control variables

The research design presupposes distraction to be at the origin of a potential difference between the two settings. The test collected the following information:

Multitasking[4]:

– existence of other open programs or browser windows during the test period;

– frequency of looking at these programs;

– estimated distraction level due to these programs.

Contacts:

– existence of contacts by phone, SMS or face-to-face during the test;

– frequency of being contacted;

– estimated distraction level due to contacts.

Technical problems:

– existence of technical problems during the test time;

– brief descriptions of any problems in the participant's own wording.

There are additional factors which might be responsible for a difference, too, or might even be the real cause. For this experiment, a catalog of possible factors was collected based on findings in different disciplines. For example, psychology expects mood and levels of attention to play

[4] An alternative form of collecting data on distraction could have been the use of spyware, which then would have raised significant ethical concerns.

a significant role in behavior studies (e.g. Ruder & Bless, 2003); social sciences expect demographics like age, gender, academic specialization or education level to have an effect (for example Diekmann, 2005) and library and information science expects the information use environment such as the kind of internet connection to be important.

This experiment collected or subsequently computed approximately 90 different variables to provide a comprehensive picture of what happened during the test situation and to gain knowledge about factors influencing the data in a natural environment. In general, studies are unable to collect such a complete catalog of potential influences, because the test size itself grows quickly and the data analysis becomes very complicated. While the focus here was on time and distraction, as wide a range of variables as possible were collected in order to determine which ones are crucial. Appendix 1 offers a complete list of all variables that were collected or computed. They are grouped in variables on time scores, on task completion, on judgments and control variables. At least in theory, each group could influence one of the other groups and therefore each variable within a group could influence another variable within a second group.

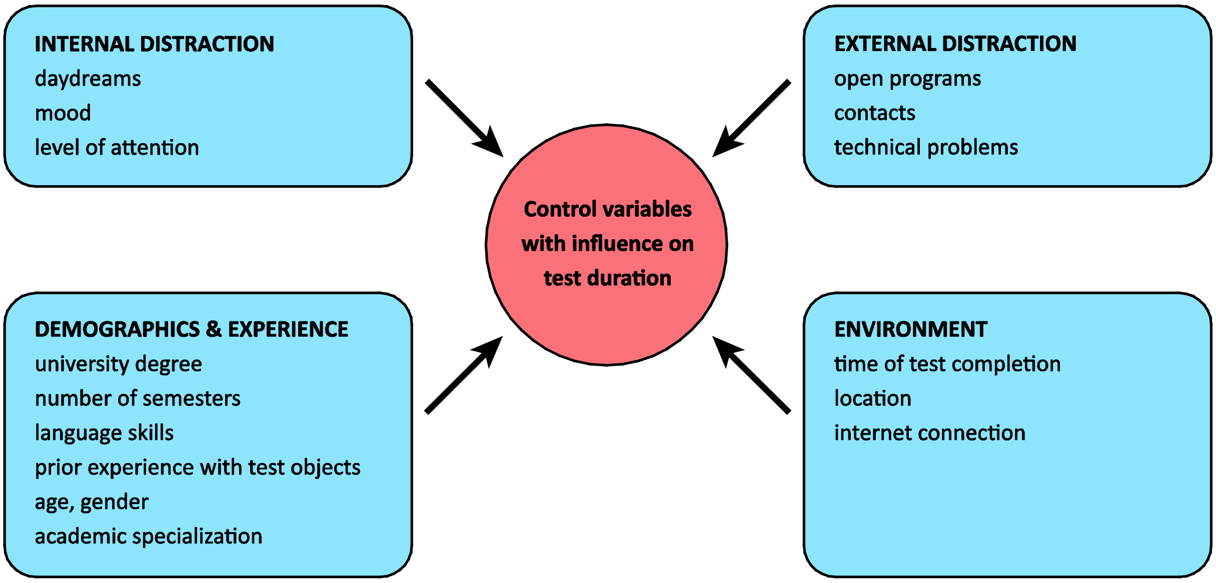


Figure 2. Overview of control variables potentially influencing *test duration*.

Figure 2 shows a grouped overview of control variables, which potentially influence *test duration.* These variables were collected after the tasks. Within the demographics and experience group, participants provided information about their age, gender, the academic specialization they were currently studying, their highest university degree and their current number of semesters studied. Participants were asked if German was their native language and they had to estimate their English skills by describing in how well they can read English-language texts and speak English. Participants were also asked if they were already familiar with one or more of the digital libraries.

The group of factors on internal distraction tried to estimate the participants' mood and their level of attention on the test. A participant in a bad mood tends to rate a website more negatively than a happy participant. Daydreams like thinking about tomorrow's shopping list or the exam next week can affect the *test duration* and the seriousness of task completion. Data on daydreams was collected as well.

In addition to the external distractions described above, additional information on the environment was collected. Participants in the natural environment could indicate the time and place that they executed the test and these decisions might have influenced the data. The hour of test completion as well as the detailed location was collected, together with information about the kind of internet connection. The environment holds without doubt many more additional factors, which might have affected the *test duration* like browser type or computer system, but could not be gathered on top of everything else.

4.6 Limitations

The research design had several limitations. Information about distraction in the natural environment was collected using survey techniques, which means this information was claimed behavior and might not be wholly true. Participants provided details on the number of open applications and the names of the concrete applications they were using. Participants were not reluctant to provide that information. One participant even reported an illegal file sharing website that was running during the test. Nonetheless, it was impossible to gain an absolutely accurate picture of the real situation with survey techniques. This was a limitation in the sense of an underestimation rather than a falsification. The research design expected participants to offer an overly positive picture of their real situation during the test instead of exaggerating their level of distraction. Hence, the findings about distraction in this work might be a conservative estimate.

The likelihood of a type 2 error, meaning the likelihood that the null hypothesis was accepted when it should be rejected, was another limitation of this study. Where possible, the analysis considered the power of the test to determine whether a larger sample might change the result. This limitation applies to some tests where smaller samples had to be used, for example tests that measured the influence of external distractions like contacts. The group size there was limited to participants in the natural environment and then to a subgroup of those who were contacted. Recruiting a much bigger sample in the natural environment in order to reduce the risk of a type 2 error would have lead to a disproportionally high number of participants in the laboratory.

The choice of software was another limitation. The software *Loop11* was chosen, because it offered the possibility to use an affordable asynchronous remote usability test tool that had minimal features and collected only quantitative data during the tasks. *Loop11* provided the output that most products offer such as: time spent on the tasks, time spent on the test, number of page views, and successful task completions. Other software products handle some things differently or perhaps even better. *UserZoom*, for example, allows participants to write the task answer directly within the same task window. *Loop11* did not provide that option, and because of that restriction, this research design chooses not to require participants to give answers at all. In *UserZoom*, “success” is measured by the answer provided in the text field and not by the success URL as in *Loop11*. This was certainly a better option for digital library user studies, but *UserZoom* currently starts its pricing at about a $1000 per month and *Loop11* offers a $350 one-time-payment price for the completion of a whole test. Many libraries will opt for *Loop11* or similar less expensive products, because they lack the financial resources to buy more costly and more sophisticated products. The limitations of this product choice can also offer an advantage: others researchers can learn what data they can expect from a more affordable product.

This study used tasks that had one clear endpoint and one clear success URL. It was expected that participants could either find that URL or that they did not find the requested information. The tasks were as precise as possible and asked users to locate a specific piece of information like the number of pages a book had. This research was aware that task complexity affects search behavior and that more complex tasks might have lead to different results. However, the tests did not include open-ended tasks, because it would have made a comparison between the two groups impossible. While the use of tasks with clear endpoints is a limitation, it has the benefit of allowing other future researchers to examine the output and to decide whether the data from this cost-effective system would be useful for their own digital library user studies.

Loop11's definition of task completion held two other potential limitations. *Loop11* evaluated unsuccessful task completion (that is everything except the success URL) as a task "failure" when participants clicked on "task complete" (and not on "abandon task"). This notion of "failure" can be problematic, because it implies a negative assessment of a user's behavior, which might be misleading. One could argue that participants who "failed" did so deliberately, because the participants in this test did not know explicitly that each click was tracked, so they might have thought that the researcher would not see what they did. But of course this understanding of failure is misleading, because it presupposes a deliberate intention to dissimulate. There could be a multitude of reasons why a participant clicked on "task complete" without ending on the success URL. Chapter 7 will discuss this understanding of a "failure" in more detail.

Loop11's definition of successful task completion has a second limitation. Their way of measuring task completion expected participants to stop exactly at the requested information—and not before or afterwards. This expected behavior excluded a phenomenon which Herbert Simon called "satisficing". It describes a user's behavior in decision-making in which users do not aspire the perfect solution, but opt for one that is good-enough. This behavior was confirmed for digital library users (e.g. Agosto, 2002). This could mean that also in test situations with clear endpoints users satisfice even if they ought to complete the tasks. If users choose to satisfice in tasks with specific required endpoints, then *Loop11's* mechanism for measuring task completions becomes problematic.

4.7 Summary

Eight recruiters asked students of all kinds during a one week period to participate in the test. Recruiters strictly alternated between participants for laboratory and natural environment. The participants did not know that there were two different settings and they had no choice between the two settings. Participants had to complete brief research tasks in five digital libraries with different languages and differences in difficulty. If participants found the requested information, they could click on "task complete", or if not, on "abandon the task". Participants could also click on "task complete" without the relevant information. *Loop11* then marks their behavior as a task "failure".

In total, approximately 90 different variables were collected or subsequently computed in order to provide a comprehensive picture of what happened during the test situation. The software *Loop11* provided data on completion time scores and page views; a survey asked participants demographic questions and questions on the context of their usage: whether they used other applications during the test and if yes, how often; whether someone contacted them or how they rated their level of attention on the test. Participants were also asked if they experienced any technical problems with the digital library during the test period. The aim of the experiment was not to show that all factors play a significant role in an online test; instead it indicated which variables can be crucial for data collection in future studies.

The research design had several limitations. Information about distraction in the natural environment was collected using survey techniques which meant that this information was claimed behavior and might offer an overly positive picture of the real situation. A potential type 2 error was another limitation of this study, because of the comparably small samples in the statistical tests in the natural environment group in chapter 8 on the influences of distraction. The choice of the software limited the type of tasks and the measurement of successful task completion. Potential misleading results in the data output might also be due to the fact that participants tend to satisfice, even if they were supposed to complete tasks with clear endpoints.

5 Description of outliers

During data analysis, it became obvious that the data did not meet the standards of normality. Most statistical tests like the t-test or Pearson’s product-moment correlation assume normality on the distribution of scores on a dependent variable and test results must be treated very carefully when they do not meet these requirements. “An outlying observation, or ‘outlier’, is one that appears to deviate markedly from other members of the sample in which it occurs” (Grubbs, 1996, p. 1). Outliers indicate discrepancies in a data set and, in general, researchers wish to get rid of them as soon as possible.

A first boxplot (figure 3) shows the variable *test duration,* which was measured by the duration in seconds to complete (successfully or not) all tasks and all questions. The boxplot illustrates that there were two extreme points in the natural environmental setting marked with asterisks (Participants 20 and 42) and a few other outliers in both settings.

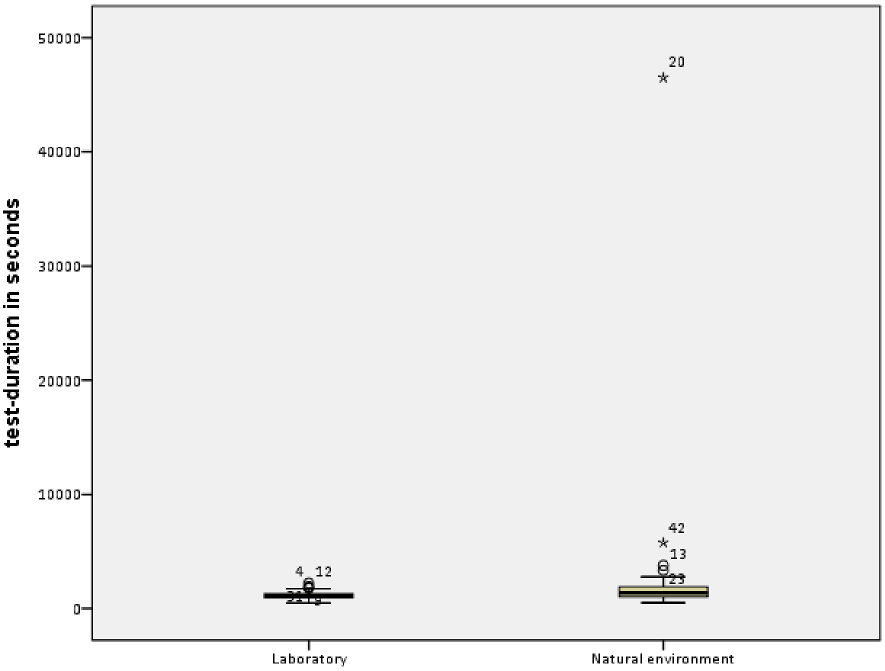


Figure 3. Boxplot of time scores on the variable *test duration* for all participants.

Recent publications by Bolton & David (2002), Hodge & Austin (2004) or Rousseau et al. (2011) suggest a variety of approaches to dealing with outliers, some of which are pursued below. However, the primary aim of this chapter is not to demonstrate what to do with outliers, but to provide a description of circumstances that make participants become outlying observations in a natural environment setting.

Grubbs' work (1996), which is still the reference for research on outliers, defines two variations of outliers: an "outlying observation may be the result of gross deviation from prescribed experimental procedure or an error in calculating or recording the numerical value" (p. 1). In accordance with variation one, the data was checked for input errors, but none could be found.

Grubbs describes the second variation as follows: "An outlying observation may be merely an extreme manifestation of the random variability inherent in the data. If this is true, the values should be retained and processed in the same manner as the other observations in the sample" (Grubbs, 1996, p. 1). This means that an outlier might be only an extreme manifestation within the data set and not a value outside of it. Deleting this kind of outlier means concealing and thus distorting the real distribution. Grubbs' quote contains another important point about the variability inherent in the data. If the detected outliers are not errors, then data's variability in an online user test of digital libraries may be higher than expected. That is, if the outliers are not just errors, the outlier's score must be treated as genuine.

The experimental data shown in figure 3 are not the first where a study struggled with how to understand outliers. Johansen et al. (2011) report—without further details—on the outliers of a usability study that "401 samples out of the 57,600 were considered outliers and removed from the analysis". Nielsen (2006) analyzed a huge quantitative data set for differences in behavior between males and females and discovered that of "1,520 cases, eighty-seven were outliers with exceedingly slow task times. This means that 6% of users are slow outliers". Instead of deleting the outliers, Nielsen (2006) examined them more closely and concludes that "slow outliers are caused by bad luck rather than by a persistent property of the users in question" (online). What he calls "bad luck" is a test situation, in which a participant might overlook a link, or got on the wrong track and was unable to find the answer. Nielsen also makes clear that these outliers are unwelcome for statistical tests, but are too many to be thrown out of the analysis.

The subsequent section examines whether the outliers in the data set collected for the experiment can be explained as "bad luck" or if there are additional factors that have an influence on the completion time measured by the variable *test duration*. The two extreme points, participants 20 and 42, do have a strong influence on the data and are a good starting point. The mean of the variable *test duration* calculated with both extreme outliers is 1,925 seconds (~32 minutes) and without the two participants, it is 1,327 seconds (~22 minutes). The standard deviation for the whole group is *SD* = 5,009 seconds and without the two extremes, it is *SD* = 587 seconds. The last number meets the expected standard deviation in usability studies, which should be at 52% of the mean value, according to Nielsen (2006).

In the following comparison of single participants' time scores, these scores are compared to the mean of the group without the two extreme outliers (that is *M* = 1,327 seconds). It is evident that the two extreme points might be 'worthy' participants, but probably have to be thrown out before running any tests. Before this dramatic step, it is worthwhile to look more closely at the two individual participants and their test behavior. What makes them an outlier? Figure 4 shows the boxplots of the variables *time on the task in (Perseus, SSOAR, …),* which measured the time spent on the task in each of the digital libraries (natural environment and laboratory).

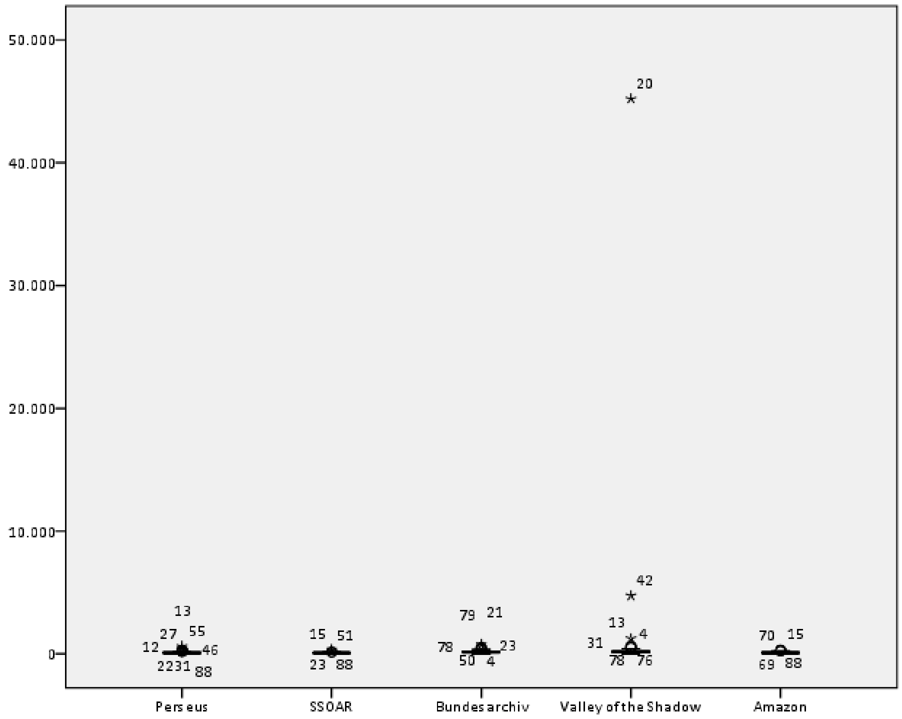


Figure 4. Time spent in seconds on each task in the five digital libraries including outliers.

The figure illustrates that both participants are only marked as outliers for the task in the digital library *Valley of the Shadow*. This means, they have met the normal distribution of this sample, except of this particular task. Following Nielsen's suggestion, they must have had "bad luck". In order to investigate what else might have happened during the test, it makes sense to use the additional information gathered in the test.

Participant 20 was a young female doing the test at home. She needed 46,488 seconds on the whole test, which means about 12 hours. Obviously, this number seems to be unrealistic. However, it is no error in *SPSS* or a data export error. It's possible, of course, that an error on the software side occurred whilst logging that particular participant.

The young female needed in average 4.5 page views to complete a task, which is even faster than the mean 5.9 page views of the whole group. She completed all tasks as fast as the average participant and completed all tasks but *Valley of the Shadow* successfully. Only for the task in this specific digital library she needed 45,213 seconds (compared to the mean of 210 seconds). She clicked "task complete" without the right result after four page views (the mean score for *Valley of the Shadow* was 11.2 page views).

Additional information reveals that she was using her iPhone to complete the test. She admits that she had several programs open during the test and that she looked at least five times at the programs. She clearly is a multi-tasker. She also admits that apart from being distracted by other applications, she was contacted (either by phone, SMS, or in person) during the test. She admits that she was disturbed by these contacts, but the distraction was not excessive. If the strange number of 12 hours spent on the task in *Valley of the Shadow* is not an error, it also might be that the participant was disturbed during the test and then forgot about the test for several hours and discovered the open test window later and decided to go on. This might be an unusual test behavior, but not an unrealistic one.

Participant 42 was also a female who did the test at home. She needed 5,765 seconds, which is above an hour and a half for the whole test. This also seems a bit exaggerated, since the mean score to complete the test was 22 minutes. As in the first case, this participant did not need many page views to complete tasks and was in general successful (her average page views per task was 6.8 and the mean score of the whole group was 5.9 page views). Again, she needed most of the time on the task in *Valley of the Shadow* (4,738 seconds compared to the mean score of 210 seconds), but needed also many more page views (24 page views compared to the mean of 5 page views). The data states that she abandoned the task. In her comments she wrote that she tried to complete the task, but could not find the requested document. It seems wrong to throw out a participant who tried hard to solve a task but failed. She obviously falls into Nielsen's category of "bad luck". She had one other program open during the test, but claims not to have been distracted by it. She admits a very strong disturbance by someone who contacted her during the test. It is hard to tell without further ethnographic or interview data whether mere "bad luck" made her be so slow or whether the disturbance had a significant effect on her time score.

Participant 42 volunteered additional information: when asked if any technical problem occurred during the test, she stated that the digital library *SSOAR* did not load. This was surprising news, since *SSOAR* is a trusted repository that should be accessible all the times and should not have loading problems. In the case of participant 42, this additional information about server problems could help in understanding her behavior in this particular digital library: she needed 136 seconds (*M* = 73 seconds), but only one page view (*M* = 3.6 page views). She clicked on "task complete" with all other tasks, even with the more complicated task in the *Bundesarchiv*, but abandoned *SSOAR* which was one of the easiest tasks. Obviously, she was inculpable, because due to the server outage she was unable to do the task. It would be hard to explain strange numbers in online tests without intentionally gathering a minimum of context information, and information about technical problems appears to be a particularly valuable variable to collect.

Deciding what to do with participant 42 is difficult. She could not complete a task and needed much time, because she had several forms of "bad luck": she was disturbed and one of the digital libraries did not load, so that she was unable to complete the task. It is tempting to treat her as an outlier that would be better to throw away to improve the normality of the distribution, but in fact she also displays important features of the variety of conditions for remote usage.

A standard outcome of the behavior of outliers 20 and 42 is to throw them out. But what happens when they are excluded from the sample? Figure 5 and figure 6 show that the number of outliers increases without the two extremes. This is not surprising, since the elimination of the two extreme points changes the mean value and now, in relation to the mean, other scores become outliers. Figure 5 shows the time spent on each task, depending on the setting. It is clearly visible that outliers occur in both settings—in the controlled laboratory environment and the natural environment—with a higher percentage of extreme outliers in the latter. It is apparent that there are too many extreme points (marked by asterisks) to suggest a normal distribution. It also becomes clear that the overall variable *test duration* masks the real test situation: rarely is one participant an outlier in several tasks.

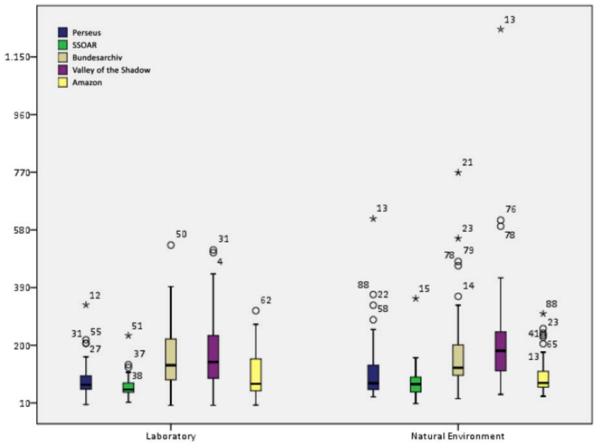


Figure 5. Time spent on each task in the five digital libraries without participants 20 and 42.

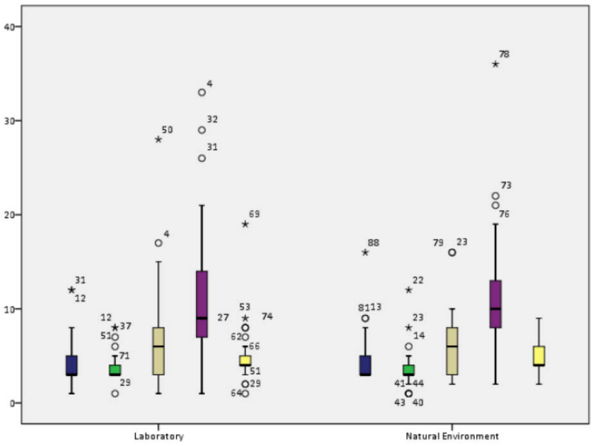


Figure 6. Number of page views on each task without participants 20 and 42.

Participant 51 in the laboratory is an outlier on the task in the digital library *SSOAR*, but not in any other tasks. On the other hand, participant 12 is only an outlier on the task in *Perseus*. This result is confirmed by Nielsen's findings: "The most seemingly obvious explanation for these outliers is simply that a few people are almost incompetent at using the Web, and they'll show up as slow outliers every time they test something. But this hypothesis is false. Once we recruit people for a study, we ask them to do multiple things, so we know how the slow outliers perform on several other tasks. In general, the same users who were extremely slow on some tasks were fast on other tasks" (Nielsen, 2006, online).

Figure 6 shows the same tasks in the five different digital libraries, but gives the variable *page views on the task (Perseus, SSOAR, …)*. Again, many outliers pop-up for both settings, with extreme points as well. However, the outliers in figure 5 are different from those in figure 6. For example, participant 21 was very slow, but did not need many page views to complete the tasks.

At this stage, it is necessary to describe other types of outliers that occur and to consider what makes these outliers become outliers. Is it higher variability on the distribution than expected, or is it only distraction that produces outliers? If the latter, would it be enough to add a pause button to avoid outliers? Participant 13, also a female, offers a relevant example. She did the test at home, and for the whole test she needed 3,341 seconds—close to an hour. Her average number of page views to complete a task were only 6.3 (compared to *M* = 5.9 page views). This means she was slow, but did not need many more clicks to complete the tasks than the average. She completed all German digital libraries (*SSOAR, Bundesarchiv* and *Amazon*) successfully, but abandoned both of the English language tasks, *Perseus* and *Valley of the Shadow*. She needed 617 seconds to finally abandon the task in *Perseus* (compared to the mean with 210 seconds) as well as many more page views (9 compared to the mean 4.3). This indicates that she had actively searched for a long time. The same was true for her behavior in *Valley of the Shadow*. Participant 13 said that she had a single program open during the test and that she looked at it three times, but she said that she did not feel distracted by it. She was also contacted once, but again felt little distraction. Compared to the other two outliers described above, she does not have a single obvious break that can be explained by a disturbance. At the end of the test, participants were asked to estimate their German and English skills and she rates her knowledge of English as very low: “I can only read it with trouble”. A lack of language skills is unsurprisingly another factor that influences the distribution of time scores and page views.

A completely different type of outlier is illustrated by participant 4: a young man doing the test in the laboratory. With 1,869 seconds to complete the whole test, he is above the mean value of 1,327 seconds and he needed nearly double as many page views as the average (10.5 page views). He also mentions technical problems with the *Bundesarchiv*, which might explain the relative high time and page view score in that particular digital library. Nonetheless he was one of the few participants who were able to complete all tasks successfully. He says he is a PhD student. In other words, participant 4 was neither distracted nor had problems completing the tasks. He was a very well intentioned participant who wanted to complete the test in the best possible way. Is participant 4 an outlier, because his behavior is too perfect to represent real retrieval behavior or does he only represent another end of the variability of data? He is an outlier in the sense of being too perfect, but excluding him would clearly be problematic.

The previous boxplots suggest that the data produces predominantly outliers on the upper end (that is, slow participants). A transformation into a standardized distribution by taking the logarithm of the time scores represents a standard method for solving the problem. A logarithmic transformation has “the consequence of bringing the tail involving slower latencies closer to the center of the distribution and making the mean a more accurate reflection of the central tendency of distribution”

(Fazio, 1990, p. 85). The expectation is that the new logarithmic values will generate a standardized distribution. All statistical tests could then be performed—even those relating to the natural environment. To test this, a new set of boxplots was generated on the logarithmic values of the variables *time spent on the task in (Perseus, SSOAR, …)* (figure 7) and on the logarithmic values of the variables *page views on the task in …* (figure 8), respectively figure 5 and figure 6 in logarithmic scale.

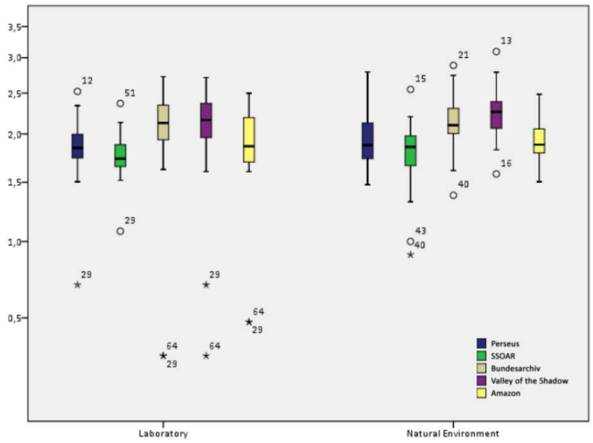


Figure 7. Logarithm: Time spent on each task without participants 20 and 42.

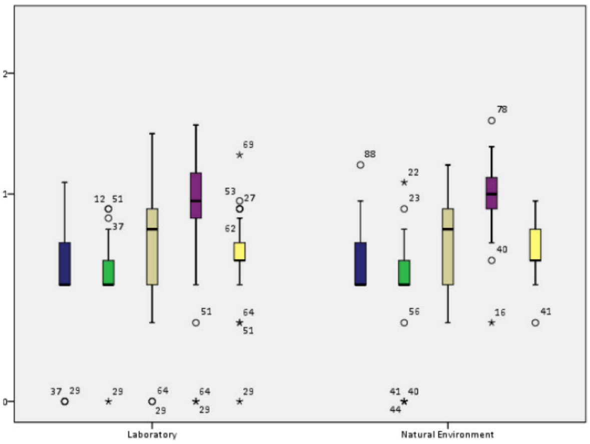


Figure 8. Logarithm: Number of page views spent on each task without participants 20 and 42.

As expected, there are few outliers in the natural environment (figure 7), but suddenly two outliers appear in the laboratory setting that have not been clearly visible before. These are “fast” outliers, ones at the lower end of the time scale. One could assume at this point that the new “fast” outliers behave in the same way as the “slow” outliers did: that is that they were only outliers in one task and are not general outliers for all tasks. However, the boxplots of figure 7 and figure 8 show that the “fast” outliers follow another pattern. “Fast” outliers are outliers that appear in nearly every digital library and they are outliers both in the completion time and the number of page views. This means that these participants have neither spent much time on the individual tasks nor have they needed many clicks for the task.

Participant 40, a male participant at home, needed only one page view in the digital library *SSOAR* and spent only 8 seconds on the task. His comment reveals that *SSOAR* did not load. This explains the one outlying score of this participant. Participant 41 and 43, both females doing the test at home on the same day, but some hours later, showed a similar behavior. Both report that *SSOAR* had problems loading and therefore appear as an outlier for that particular task.

Participant 41 appears a second time as an outlier for the task in *Amazon* (figure 8). An in-depth analysis of her test situation shows that after her technical problems with *SSOAR*, she finished the task in the *Bundesarchiv* without problems. She struggled in *Valley of the Shadow* and abandoned the task, after she could not find the right result. In *Amazon*, where she becomes an outlier on the page view score again, she makes only two clicks, without actually starting a search. She fails, but still spends 205 seconds on *Amazon* (M = 105 seconds) which is the highest score for a task she spent during the whole test. In her comments, she says that she had *Skype*, *Facebook* and an email program open, and looked at them once during the test, but did not feel distracted. It looks as if this was a false impression on her side. It seems likely that after her difficulties with *Valley of the Shadow*, she lost interest in the test and moved her full attention to other things.

In the laboratory setting two outliers appear several times in both boxplots: on the one for the five variables *time on the task in (Perseus, SSOAR, …)* and on the one for the variables *page views on the task in …* Participant 64 is a female whose native language was not German. She was the only one that said that she did not concentrate at all on the test (of course, this question came after the test and participants were assured that their answers had no influence on the reward). Apart from *SSOAR*, she failed all tasks. At the end, she abandoned the task in *Amazon*, maybe because she was no longer in the mood for tests. After her successful participation in *SSOAR*, she spent only between 2 and 5 seconds on each task; she finished all tasks in 119 seconds (*M* = 662 seconds) and took only 478 seconds on the whole tests, including the questions (*M* = 1327 seconds). It is interesting that she always clicked "task complete", even when she must have known that she had not found the right result.

Participant 29, a male, was the fastest participant in the test with 193 seconds on the whole test (*M* = 1327 seconds). Obviously, he could not have left the laboratory after 3 minutes, so it is to be supposed that after the test he continued on a personal task without being noticed by the recruiters in the laboratory. He had an average page view rate of 1 click per task, which means he must have started the task and immediately clicked on "task complete". Like participant 64, he clicked on "task complete" even without having completed the task. He also had an inconsistency within the answers of his test questions.

A last outlier in the laboratory was female participant 51. She follows a similar pattern to participant 64, but is smarter. She completed the task in *Valley of the Shadow* in 66 seconds and 2 page views. Anyone who has already used this particular digital library knows that this is simply impossible. A minimum of five page views was necessary to complete the task. However, her second page view was the requested page. It can only be speculated that she used a search engine like *Google* to find the right result.

It is interesting that all three participants, who in some sense cheated on the tasks, turned up in the laboratory under controlled circumstances. None of the participants in the natural environment obviously speeded through the test without actually doing it or using external help tools. This circumstance is remarkable, because it would have been much easier to cheat in the natural environment. It is also good news for tests in online natural environment settings.

Based on the previous analysis, outliers can be grouped into seven variations:

- The “speeder”

  This kind of outlier could theoretically turn up in either a laboratory or a natural environment. Outliers in that group can be identified as “fast” outliers by a boxplot of logarithmized values. Speeders are repetitive outliers on the variables *time spent on the task …* and the variable *page views on the task ...* Note that the speeder can be different than what *Loop11* describes as “failure”.

- The “scrupulous”

  This kind of outlier is mirrored by participants that are overly scrupulous in the test situation and try to accomplish tasks in the best possible way. Scrupulous outliers can be identified by a single boxplot of the variable *test duration*. Indicators are slower time scores than the mean and successful task completions for most tasks. This outlier can turn up in either a laboratory or a natural environment.

- The “unlucky”

  This outlier, as identified by Nielsen (2006), mirrors a test situation in which the participant gets lost. It can turn up in either a laboratory or a natural environment. This kind of outlier can be identified by a clustered boxplot with a single high number on one of the variables *page views on the task in …* and a high number on the corresponding variable *time on the task in …*, combined with an average time number on other tasks.

- The "multi-tasker"

  This kind of outlier appears in the natural environment. This outlier switches between tasks during the test time. The effect is an overall strong influence on the variable *test duration*. It can be identified by a single boxplot of the variable *test duration*. The number of page views of multi-tasker is close to the mean; and multi-tasker have a slightly less successful task performance.

- The "break-taker"

  This kind of outlier is marked by a strong external disturbance during the test, which requires the full attention of the participant. The break-taker can be identified by a clustered boxplot of the variables *time on the task in …* The break-taker can be identified by an exponentially slower time score than the mean score. The outlier's page view score meets the mean. It appears in the natural environment.

- The "inculpable"

  This outlier meets with an external disturbing factor on which the participant has no influence. This can be, for example, a server problem. This outlier can only be identified by additional context information. It mostly appears in natural environment settings; participants in a laboratory could orally inform the researchers.

- The "handicapped"

  This outlier appears if participants do not have required competences for specific tasks, for example the language knowledge to search in English sites or a particular cultural or historical knowledge. This outlier can only be identified by adequate context information. This outlier can turn up in either a laboratory or a natural environment.

The previous description makes clear that outliers are not only caused by "bad luck". It also became obvious that distraction is an important (disturbing) factor in asynchronous remote usability tests. The analysis showed that not only distractions such as multitasking or a direct contact affected test data. The "inculpable" outlier meets with an external disturbing factor on which the participant has no influence. This can be, for example, a server outage. It is an inherit problem of online tests and context information is the only way to interpret that kind of outlying observation.

There are several human factors that have an influence on the data too, and are not caused by the natural environment. The “handicapped” outlier appears if participants do not have required competences for specific tasks, for example the language knowledge to search in English sites or a particular cultural or historical knowledge. While these factors also apply to non-remote test situations, they are all the more important in an asynchronous remote usability test setting in a natural environment, because researchers could not see when participants are having a problem, as they could in a laboratory setting.

The “scrupulous” outlier is a known phenomenon of test situations. It is mirrored by participants that are overly scrupulous in the test situation and try to accomplish tasks in the best possible way. The “speeder”, on the other hand, manipulates explicitly a test situation. It might have been a chance circumstance that both types of outliers appeared only in the laboratory, and that these human factors seemed to play an insignificant role in remote natural environments. This sample was too small to establish that as true or not, but it raises the interesting possibility that online tests in natural environments could allow a more adequate collection of real information behavior than laboratories.

The description can give no general advice what to do with the outliers. In the end, it depends on the research question that the data should answer. In order to answer to the original research question of this dissertation—involving a statistical comparison between two settings—there is probably no other possible way than to exclude the extreme points which were described in this chapter. These are participants 20 and 42, 13, 29, 40, 41, 43, 51 and participant 64. An exception will be made for participant 4, the “scrupulous” user, since that type of behavior is not unrealistic in everyday use situations.

6 Formal data description

6.1 Demographics

After excluding the outliers (chapter 5), 75 participants took part in the experiment with 38 participants in the laboratory and 37 participants in the natural environment. A bit over half of the participants were female (64% females and 36% males). Participants had an average age of 26. Most participants' native language was German; 17.3% were non-native speakers. The section 6.3 provides an overview table on demographics and levels of distraction. The sample matches the student population at the Humboldt-Universität zu Berlin who have an average age of 23 years (Ramm et al., 2011), with a proportion of 58% females and a 17% share of foreign nationals (Humboldt-Universität zu Berlin, 2012). The university student population differs slightly from the library lobby population, because it does not include library visitors from other universities. The exact distribution of the library population is unknown at this point, because precise data are non-existent.

In the laboratory, 57.9% participants were females and 42.1% were male participants with an average age of 25.6 years. In the natural environment, 70.3% participants were females and 29.7% were male participants, with an average age of 26.3 years. The distribution of native speakers and non-native speakers was close in both settings with 21.1% non-native German speakers in the laboratory and 13.5% in the natural environment.

More participants in the laboratory judged their level of English as medium or low (21%) than in the natural environment (8.1%). In fact, an imbalance between the two samples might be the reason for that difference and possible implications will be verified in chapter 8 (findings on control variables). An alternative explanation draws on potential effects on participant's behavior in a laboratory which might lead to a more negative assessment of one's own capacities in order to make the recorded retrieval behavior on the English sites look more positive. Regardless of the reason of the imbalance, more participants in the laboratory had difficulties with German and with English as well. Lower language skills lead to an increase in usage time, because participants need more time to understand and complete a task. Assuming that distraction has no or only little influence on the *test duration*, participants in the laboratory should then be slower in time, because of lower language skills.

All participants in the laboratory had high speed broadband internet connection. While 13.5% in the natural environment only had mobile or slow internet connection, 16.2% participants in the natural environment used the high speed broadband connection from the university as well (for example when they participated in the test in the library lobby) and 67.6% used a high speed connection somewhere else.

The German universities changed to the Bachelor and Master system about seven years ago. There still exist a number of programs for awarding older degrees, however, and some students are still enrolled in these older programs. The majority of participants were undergraduate students. Their highest level of education was the German “Abitur” (final secondary-school examinations). About 20% were currently graduate students (they hold a Bachelor degree or “Vordiplom”). Even for the library staff, a surprising high number of 26.7% of participants already had a degree higher than graduate level (they hold a Master, “Magister”, “Diplom” or “Staatsexamen”), and additional 5.3% already possessed a doctorate. The education background was well balanced in both settings.

The majority of participants were experienced students in their 6th semester or even higher (in the laboratory 57.9%, in the natural environment 64.9%). It is a common (if undesirable) fact that undergraduates take more than six semesters to finish a Bachelor's degree, because German students pay a minimal tuition compared to the US and the financial pressure to finish is less intense. In comparison, only a fourth of the participants were university newcomers in their first or second semester (28.9% in the laboratory and 18.9% in the natural environment). The high educational level and the relatively high number of semesters at the university suggest that the experiment’s participants should in theory have rather advanced existing task related knowledge about how to find documents for assignments.

German students rarely study only one academic subject, but rather pursue a combination of majors and minors. In the following description the first, second, or third subject studied counted equally, because for the analysis, it did not matter whether a student studied history as a major and linguistics as a minor. The important information is that this participant had a background in both academic specializations. Participants were students from the Humboldt-Universität zu Berlin and also from other Berlin and Brandenburg universities. Participants listed 63 different majors and minors that they were studying. For ease of reading and analyses, these 63 subjects were grouped into three orientations (see appendix 5 for a detailed list of individual academic specialization with their original German name as provided by the participants). The three areas point to the degree of text orientation, technical background or knowledge of test construction.

Figure 9 shows the distribution of academic specializations in both settings. The three groups represent different kinds of academic specializations. Starting with text oriented subjects, the amount of experience with texts and therefore experience with libraries and databases slowly decreases to the second group with economics and math studies and to an increase in knowledge of how to deal with numbers. Participants in the last subject area have experience with test designs and potentially show a distinctive test behavior. They have experience with texts and technique. The distribution is similar between the settings and shows that in both settings most participants had experience with texts.

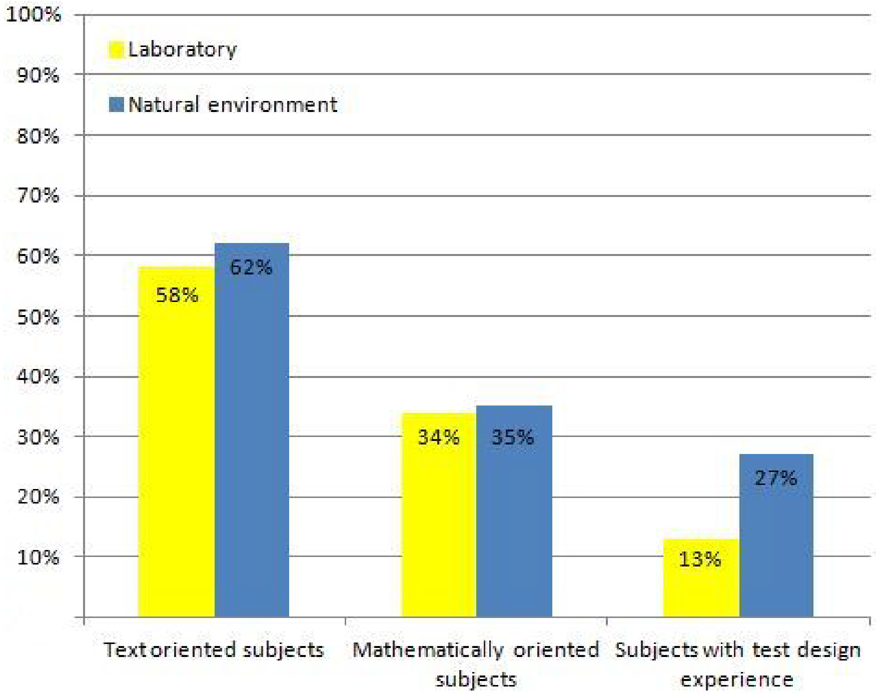


Figure 9. Distribution of academic specializations in both settings.

Mood plays a crucial role in test behavior. Participants in both settings showed a variety of moods. The average mood was similar in both settings with participants rating themselves 4.9 in the laboratory setting and 5.1 in the natural environment (1 means a negative, 8 a very positive mood). In general, participants were in a good temper. Figure 10 illustrates the distribution of mood in both settings: the deeper the orange, the less positive was the mood. Some points on the Likert scale were never chosen by participants, so figure 10 shows only six categories.

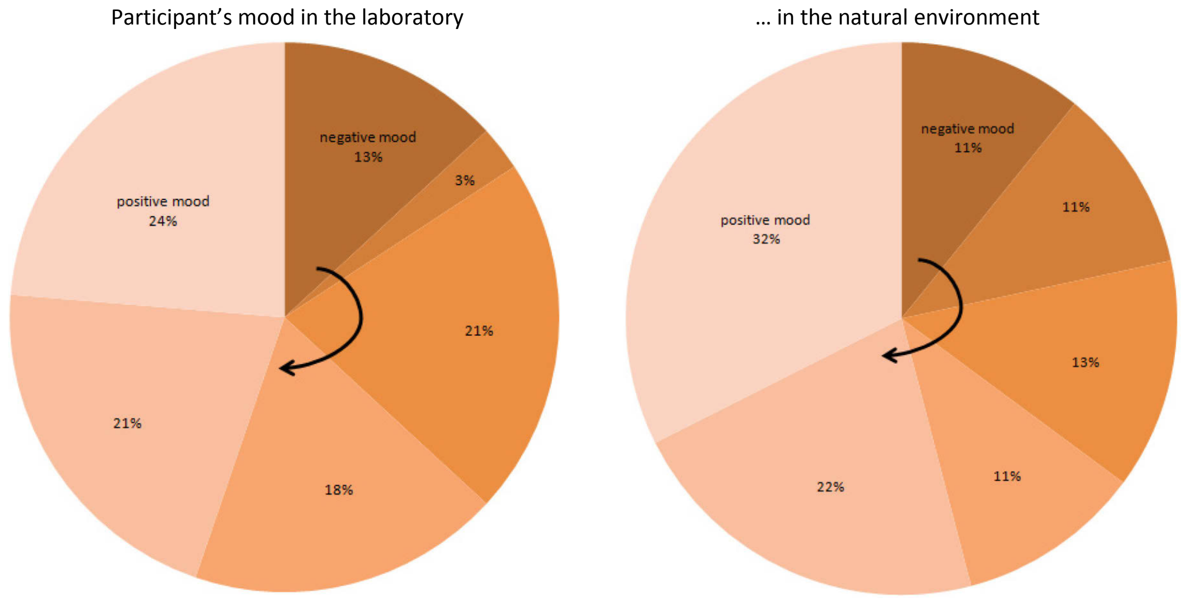


Figure 10. Distribution of mood in both settings.

6.2 Level of distraction

A principal element in the natural environment is the existence of distraction. The experiment was designed to give evidence that this is also true for test situations in the natural environment. After task completion, participants were asked to describe their behavior and contextual environment. Data on open programs and multitasking, on disturbances like contacts in person, via telephone or messaging, or on distractions like daydreams were part of the survey. Some questions only applied to the natural environment, since in the laboratory external distractions were reduced to a minimum.

A large majority—64.9% of participants in the natural environment—admitted that they had another program or application open during the test period. This could have been another program like a chat or an email program, but also other browser windows. Multitasking was in general limited to one, two or three open applications in addition to the test window. Of the 64.9% of participants with open programs had 40.9% only one program open, 31.8% had two or three programs open. 18.1% had even four to six programs simultaneously open—in addition to the test window—and a small group of 9% of participants admitted that they had more than 16 programs or browser windows open. Given those numbers, it was possible to validate the central research hypothesis of the existence of potential distraction in remote test settings in natural environments.

Many participants had different kinds of programs open. It is not surprising that email was the most popular program. 83.3% of the participants, who had other programs open, had their email program open, too. If incoming email notification is set up, distraction is preprogrammed. Other strong distractions by proactive programs were less in use. 29.2% had chat programs open and 16.7% had *Facebook* open. Browser windows were a favorite distraction, too. 45.8% of the participants had other browser tabs or windows open. 20.8% had some form of music program open that played music on the computer and 12.5% had a variety of other programs such as dictionaries open.

While these numbers indicate a strong and continuous distraction by multitasking, participants estimated their own distraction as marginal. Only 11.2% of the participants said that they actually have looked at the other programs during the test time. All other participants said that these programs were open, but that they have never looked at them. Consequently, 64% admitted no distraction at all by the open programs. 28% felt a light distraction due to the open programs. Light means they rated their distraction as 2 or 3 on an 8-point Likert scale. Only two participants admitted a rather strong distraction (that means 8% rated their distraction as 4 or 6). Not a single participant estimated the distraction as intense. Having other programs open does not automatically mean that participants are doing multitasking. Participants experienced the sheer fact of having other programs open not as a distraction. The outlier's analysis in chapter 5 showed that this impression is sometimes erroneous.

Multitasking is a newer phenomenon that increased in popularity with broadband internet connection, social networks, and ubiquitous computing. Researchers are aware that multitasking is an elemental part of the users' natural environment. However there is a much older form of distraction in the natural environment: human contacts. In their natural environment, external persons may address the participant face-to-face, by phone, or by sending a short message. Participants with family at home are prone to be addressed, as well as participants in public places like the outlier "participant 20" (see chapter 5).

The postulation of disturbances due to contacts turned out to be true. 27% of all participants in the natural environment were contacted either in person or by phone during the test period. A contact requires a mental shift into a different kind of activity (see chapter 2.2.2) and is a strong disturbance within a process. A large majority (90%) of these participants had been contacted once or twice and only 10% were contacted more frequently. Even if the number of interruptions was minimal, each interruption has a potential influence on user's behavior. For example, a participant takes part in the test and is interrupted by a friend. The friend invites the participant to have a cup of coffee. Even if the participant decides to finish the test first, the level of attention on the test is likely reduced or lost after the interruption.

80% of participants who were contacted estimated their resulting distraction as non-existent or marginal. Only two participants admitted a distraction, no participant experienced the contact as a strong distraction. Having demonstrated the influence of distractions—especially of contacts—on the outliers and knowing about the influence of distractions on people's behavior, self-assessment of distraction appears not to be very reliable.

Technical problems are another form of distraction—even if the participant has nothing to do with its origin. In the worst case a technical problem is a server outage, but more subtle defects also have a negative influence on participants. In this experiment, participants were asked if any technical problem(s) occurred during the tasks and, if so, were asked to describe it briefly. Participants experienced long(er) loading times of websites as technical problems, because it was an uncommon situation. Participants translated this to non-working. Some participants also reported on a lack in search functionality or database problems. 39.5% of the participants in the laboratory and 32.4% in the natural environment reported technical problems. Whatever reason existed behind a reported technical problem, participants got distracted by the problem or even had to stop the test.

Technical problems are hard to control in an asynchronous remote test environment. In-depth testing and good communication about test periods are an indispensable requirement before running a test. For this experiment, a large period of pilot testing, communications with *Loop11* staff about server updates and communications to the digital libraries about the experiment could not avoid the technical problems that did occur. Since technical problems cannot be excluded in asynchronous remote settings, gathering information on them is highly recommended for data interpretation.

Daydreams are another form of distraction, which can occur during a test period either in a laboratory or in a natural environment. A daydream can be a reflection about the shopping list for tonight, the exam next week, an anecdote or similar things. Every thought that distracts the participant from its primary task—the test—can be defined as daydream. Tasks or questions can even be the origin of a subsequent daydream. A question that asks about a recent date, for example, would likely result in daydreams. In the laboratory, 52.6% of all participants stated that they had no daydreams at all during the test. Another 36.8% admitted to some daydreams (they rated their level of daydreams as 2 or 3 on an 8-point Likert scale) and 10.6% had more daydreams (4 to 6 on the scale). In the natural environment, the picture is different. Only 21.6% stated no daydreams at all. The majority with 64.8% admitted some daydreams (2 to 3 on the scale) and another 10.8% had more daydreams (4 to 6 on the scale). Neither in the laboratory nor in the natural environment admitted any of the participants a large number of daydreams.

It might be true that participants in the natural environment were more distracted by daydreams than participants in the laboratory. Another possible explanation argues again for laboratory effects (here the "subject expectancy effect") because participants might want to report the expected answer. Even if these participants have not quite reported their true feeling, their answers would have likely not have changed in a dramatic way, but have became closer to the answers in the natural environment—which means that in both settings, daydreams played a role, but were not a significant distraction.

A similar response behavior can be seen in the level of attention. Participants were asked to rate their level of attention on the test. A large majority (73.6%) in the laboratory rated their attention as high compared to 51.3% in the natural environment. 25.7% of the participants in the laboratory judged that they were more or less concentrated on the test (4 to 6 on the scale), compared to 43.2% in the natural environment. In both settings, only a minority of 2.6% (laboratory) and 5.4% (natural environment) had not concentrated at all on the test. The level of attention on the test was relatively high in both settings.

6.3 Summary

Table 3 (below) offers an overview of the demographics and levels of distractions. Several aspects only apply for the natural environment, for example additional programs were not allowed in the laboratory. 75 participants took part in the experiment using a fast internet connection while feeling in a relatively positive mood. Most participants had completed multiple semesters at a university and had an elevated educational level, which suggests that participants should, in theory, had rather advanced knowledge about how to search documents for assignments.

	laboratory	natural environment
number of participants	38	37
gender	57.9% females 42.1% males	70.3% females 29.7% males
average age	25.6	26.3
percentages of non-native German speakers	21.1%	13.5%
percentages of medium or low level English	21%	8.1%
number of semesters	57.9% more than six semesters 28.9% first or second semester	64.9% more than six semesters 18.9% first or second semester
academic subjects	strong text orientation	strong text orientation
kind of internet connection	university fast internet connection	13.5% mobile or modem 16.2% university network 67.6% fast connection somewhere else
average mood	4.9 (1 = positive; 8 = negative mood)	5.1
program open during tests (yes/no)	n/a	yes: 64.9%
number of open programs	n/a	1 program open: 40.9% 2 or 3 programs: 31.8% 4 to 6 programs: 18.1% more: 9%
kind of open program	n/a	email: 83.3% other browser windows: 45.8% chat: 29.2%; music: 20.8% *Facebook*: 16.7% other: 12.5%
looking at programs during the test	n/a	11.2% looked at open programs
percentages of estimated distraction by open programs	n/a	no distraction: 64% light distraction: 28% strong distraction: 8%
contact occurred (yes/no)	n/a	yes: 27%
frequency of contact	n/a	a single or two contacts: 90% more: 10%
percentages of estimated distraction by contacts	n/a	no or marginal distraction: 80%
reported technical problems	39.5%	32.4%
daydreams during test	no daydreams: 52.6% some daydreams: 36.8% daydreams: 10.6%	no daydreams: 21.6% some daydreams: 64.8% daydreams: 10.8%
level of attention on the test	very much: 73.6% more or less: 25.7% no concentration: 2.6%	very much: 51.3% more or less: 43.2% no concentration: 5.4%

Table 3. Overview on demographics and on the level of distraction in both settings.

Participants in the laboratory estimated their English skills below the skills from participants in the natural environment; more participants in the laboratory were non-native German speakers. Both language factors suggest that participants in the laboratory should need more time to complete the test, because they need more time to process the questions and the language on the websites.

The central research hypothesis of the existence of additional open programs in asynchronous remote test settings in natural environments could be validated. But the fact that other programs were open does not necessarily mean that participants were multi-tasking. Few participants stated that they looked at the open programs and the sheer fact of having other programs open was not experienced as a distraction. Participants in the natural environment also got contacted during the test. They estimated neither additional programs nor contacts during the test as a significant distraction. In both settings, daydreams played a minor role and participants showed a high level of attention on the test.

A number of technical problems occurred during the test in both settings, despite a maximum amount of prior testing and communication. Participants reported a variety of answers that they perceived to be a technical problem (for example slow response times). Technical problems can result in distraction or break-offs. Since it seems almost impossible to eliminate technical problems, collecting data on these is recommended.

7 Findings and discussion I: differences between the test settings

This experiment sought to examine whether the data in a participant's natural environment differs from the data from the same test in a laboratory. Building on earlier research findings, this research presupposes that there are differences in various data between a laboratory test and a natural environment using an asynchronous remote usability test. Findings and discussion are presented together. Based on earlier theorizing and empirical evidence, it was expected that between the two settings

(hypothesis 1) there is a difference in the time participants needed to complete the test (difference expected for all time variables *test duration*, *time on tasks*, *time on questions*);

(hypothesis 2) there is a difference in the variability in the time participants needed;

(hypothesis 3) there is a difference in the participants' judgment of the digital libraries;

(hypothesis 4) there is a difference in the decision-making process for participants' judgments.

(hypothesis 5) there is a difference in the number of page views participants needed;

(hypothesis 6) there is a difference in the number of successful task completions.

Within the statistic formulas, two abbreviations are used: LAB for laboratory and NE for natural environment. The analyses draw on statistical tests to compare groups and to discover relationships. All analyses were run with a pre-specified significance level of $p < .05$ (two-tailed) and with all 75 participants, if not stated otherwise. Outliers described in chapter 5 are excluded from this and all following calculations. Parametric tests were run, unless the assumption of equal variances for t-tests or the assumption of interval-scaled data for relationship testing was not met; in this case, the t-test formula was changed accordingly or non-parametric alternative tests were chosen. Following Cohen (1988), effect size estimates within .10 to .29 are categorized as small effects, within .30 to .49 as medium effects and within .50 to 1.0 as large effects.

Eight different variables were collected or subsequently computed to measure time during the test. Variable names are in italics for better recognition. Summing up the single time scores (i.e., the time to complete the task in *Perseus*, the time in *SSOAR* etc.) a new aggregate variable reflects the time spent to complete all five tasks: *time on tasks*. This measure is different than the variable *test duration*, which includes the time spent on questions as well. Subtracting *time on tasks* from the *test duration*, a new variable *time on questions* could be created, which measured the time spent on the questionnaire only. Figure 11 illustrates the three different variables.

<table>
<tr><td colspan="6"><i>test duration</i></td></tr>
<tr><td colspan="5"><i>time on tasks</i></td><td><i>time on questions</i></td></tr>
<tr><td>time to complete task in <i>Perseus</i></td><td>time to complete task in <i>SSOAR</i></td><td>time to complete task in <i>Bundesarchiv</i></td><td>time to complete task in <i>Valley of the Shadow</i></td><td>time to complete task in <i>Amazon</i></td><td>questionnaire</td></tr>
</table>

Figure 11. Variables for measuring time.

7.1 Difference in test duration

The first hypothesis of the experiment assumed that there is a difference in the time participants needed to complete a test in the laboratory and in the natural environment. Various statistical tests to compare groups were used to verify this hypothesis. The influence of distraction and control variables on time scores will be assessed in chapter 8.

Participants in the laboratory spent 1182.4 seconds (~20 minutes) on average to complete the test, which matched the officially announced test time (see chapter 4). Participants in the natural environment were slower with an average of 1491.7 seconds (~25 minutes). This difference was significant as shown below. The means on the variable *time on tasks*—that is the time participants spent on the tasks—were, in comparison, low. Participants in the laboratory needed on average 619.5 seconds (about 10 minutes) and participants in the natural environment needed 404.9 seconds (about 12 minutes) to complete the tasks. Participants in the laboratory needed on average 562.8 seconds (about 9 minutes) to complete the questionnaire. In the natural environment, participants spent an average of 786.8 seconds (about 13 minutes) on the same questions.

An independent-samples t-test was conducted to compare the *test duration* for participants in the laboratory and in the natural environment. Because the assumption of equal variances between the two groups was violated, the t-test formula was changed accordingly. There was a significant difference in scores between LAB (M = 1182.37, SD = 380.07) and NE participants (M = 1491.73, *SD* = 637.58; *t*(58.42) = -2.54, *p* = .01, two-tailed, *Cohen's d* = .59). This means, the second hypothesis could be validated: there was a statistically significant difference between the settings in *test duration*. This result agrees with earlier research findings (Czerwinski et al., 2000; Bowman et al., 2010; Kirschner & Karpinksi, 2010; Greifeneder, 2011b).

As demonstrated, the differences between the means within the data set were not as striking for all time measures. The variable *test duration* included the time spent on the five tasks (*time on tasks*) and the time spent on the questions (*time on questions*). It is now analyzed which of the underlying time scores were responsible for the difference between the settings when looking at the *test duration: time on tasks* or *time on questions*.

An independent-samples t-test was conducted to compare the time spent on the five tasks for participants in the laboratory and in the natural environment. There was no significant difference in *time on tasks* between LAB (*M* = 619.58, *SD* = 256.37) and NE participants (*M* = 704.9, *SD* = 322.32; *t*(73) = -1.27, *p* = .21, two-tailed). Given a total of 75 participants, an alpha of .05, and an expected medium effect size of .50, the test's power was large enough to detect differences if these were existent (power = .57, see Faul et al., 2009).

A second t-test showed that there was a statistically significant difference in *time on questions* between LAB (*M* = 562.79, *SD* = 152.1) and NE participants (*M* = 786.81, *SD* = 391.56; *t*(46.38) = -3.25, *p* = .002, two-tailed, *Cohen's d* = .75). That is, the variable *time on questions* influenced the more general variable *test duration*.

The differences between the settings in time scores are even more pronounced because of the fact that participants in the laboratory had weaker language skills with more non-native German speakers, and more participants with limited English (see chapter 6). Both language factors suggest that participants in the laboratory should need more time to complete the test, because they need more time to process the questions and the language on the websites. In fact, the opposite occurred.

The detected differences in *test duration*, *time on tasks* and *time on questions* have given evidence for a number of issues in natural environment test situations. There was no statistically significant evidence that the time spent on the tasks was different between the settings. Therefore, the test setting does not matter for data collection on time during task completion. Data on *test duration* was significantly different between a laboratory and a test in a participant's natural environment.

But—and this is an important finding, which was missing in earlier research attempts comparing settings—the difference does not occur in the part of a research test in which these numbers are usually gathered: namely the numbers that measure the time on task performance. The difference occurs in data from the variable *time on questions*. Or to put it more plainly: as long as participants are occupied with tasks, the setting does not matter.

Completion times on the questionnaire are of less interest to library and information science researchers as long as they do not alter the answers—and they seem not to (see below). Under these circumstances, if participants in both settings need similar amounts of time to complete the tasks, then the setting (for example the natural environment) does not matter.

Fast participants do not necessarily provide high quality answers. If participants need significantly more time to complete a questionnaire, then they might think more deeply about the answers or they might have been distracted during the questionnaire. Only the answers about distractions can help to explain the reasons why some participants took longer to complete the questionnaires. This will be further explored in chapter 8.

7.2 Difference in the variability in participants' test duration

Based on earlier findings in the pilot test (see chapter 3), the second hypothesis suggests that the variability in the time participants needed to complete a test differs in the laboratory and in a natural environment (using the same test). The hypothesis was verified by comparing extreme values.

Participants in the natural environment were not only slower on average, but the variability of the scores was larger as well. Comparing the highest and lowest scores (that is the fastest and slowest participants) in both settings revealed that the behavior differed little on the lower scores; there was no big difference between fast participants in either the laboratory or the natural environment who both completed the test unusually quickly.

In the laboratory (left side of table 4), the variability among the five highest scores on the variable *test duration* ranged within 500 seconds. In the natural environment (right side of table 4), the slowest participant took more than 1500 seconds more than the fifth slowest participant. While the 500 seconds in the laboratory relates to 30% of the fifth highest value, 1500 seconds in the natural environment relates to even 67%.

LAB	1	highest score	2227 sec	**NE**	1	highest score	3774 sec
	2		1922 sec		2		2759 sec
	3		1869 sec		3		2395 sec
	4		1782 sec		4		2294 sec
	5		1712 sec		5		2263 sec
			…				…
LAB	5		771 sec	**NE**	5		832 sec
	4		685 sec		4		780 sec
	3		655 sec		3		755 sec
	2		653 sec		2		723 sec
	1	lowest score	564 sec		1	lowest score	669 sec

Table 4. Lowest and highest time scores on the *test duration* in both settings.

Table 5 (below) shows that the variability in the variable *time on questions* between the two settings was even more pronounced than for the variable *test duration*. While both settings were relatively similar in the lower scores, the variability between the five highest scores was very pronounced. In the laboratory, the fastest participant (305 seconds) spent only seven hundred seconds less than the slowest participant at 1,038 seconds. In contrast, the fifth slowest participant in the natural environment needed more time than this with 1,054 seconds. The slowest participant in the natural environment needed 2,483 seconds which was 23 minutes more than the fifth slowest participant and 35 minutes more than the fastest participant in the natural environment.

LAB	1	highest score	1038 sec	**NE**	1	highest score	2483 sec
	2		860 sec		2		1796 sec
	3		851 sec		3		1117 sec
	4		788 sec		4		1098 sec
	5		752 sec		5		1054 sec
			…				…
LAB	5		400 sec	**NE**	5		477 sec
	4		394 sec		4		449 sec
	3		384 sec		3		425 sec
	2		362 sec		2		422 sec
	1	lowest score	305 sec		1	lowest score	371 sec

Table 5. Lowest and highest time scores on the questionnaire in both settings.

The difference in variability in the variable *test duration* between the two settings was caused by the time spent on the questionnaire (*time on questions*). The second hypothesis could be validated: there was a difference in the variability in the time participants needed to complete a test. The findings from the pilot study (Greifeneder, 2011b) could be reproduced. The analysis revealed that this was primarily in the time spent on the questions (variable *time on questions*).

A high variability is not a problem in itself. It shows that people act differently, which is a normal-life situation. However, if variability in *test duration* differs between tests settings, critical assumptions of many parametric statistical tests are violated, potentially increasing the likelihood of type 1 and type 2 errors. Researchers, who collect empirical evidence, need to find a way to equate the variability between to-be-compared settings or at least to draw on non-parametric tests that have lower power, but are less sensitive to violations of distribution. Chapter 8 will discuss which factors influence time scores and how far they can be eliminated. Of course eliminating factors in the natural environment makes the environment artificial again. Although the difference in variability on the variable *time on questions* is a threat to statistical tests, the variable can be used as an indicator for interpreting data in test situations. The variability strongly suggests that something else must have happened during the questionnaire.

7.3 Difference in the participants' judgments

Hypothesis 3 presupposes that there is a difference in the participants' judgments of the digital libraries between a laboratory and a natural environment setting. Judgments can be assessed by asking participants to rate a website or to judge the perceived difficulty of a task. In the experiment, participants completed five tasks in four different digital libraries and in one online-shop. After each task, they had to answer six questions. Four questions asked about the perceived functionality and professional appearance of the digital library (addressed in section 7.4); the remaining two questions collected information about the participants' general evaluation of the digital library and their perceived difficulty of the task. In order to reduce response bias, participants were able to skip all six questions. They could disagree (1) or agree (8) with the statements below on a Likert scale. An equal number of Likert scale options were chosen to avoid errors of central tendency and to force participants to make a decision on the direction of their judgment.

In order to measure the participants' general evaluation of the digital libraries, the participants could agree or disagree with the statement "I consider the website as very good". The term "website" was explicitly chosen instead of the more appropriate term "digital library", because "digital library" disoriented pretest and pilot test participants. Table 6 below shows the distribution of the participants' general evaluations on all five digital libraries in both settings (laboratory and natural environment). Inspection of table 6 suggests that there are differences between the settings and as well as between the tasks. For example, the percentage of agreement is higher in both settings for the digital library *Bundesarchiv* than it is for the digital library *Valley of the Shadow*.

		I disagree	**2**	**3**	**4**	**5**	**6**	**7**	**I agree**
Perseus	LAB	–	5.3%	7.9%	15.8%	23.7%	18.4%	21.1%	7.0%
	NE	2.7%	2.7%	5.4%	16.2%	16.2%	29.7%	13.5%	13.5%
SSOAR	LAB	–	2.6%	2.6%	2.6%	5.3%	31.6%	39.5%	15.8%
	NE	–	–	8.1%	5.4%	2.7%	32.4%	29.7%	21.6%
Bundesarchiv	LAB	2.6%	7.9%	21.1%	5.3%	7.9%	23.7%	10.5%	21.1%
	NE	–	–	5.4%	5.4%	16.2%	16.2%	40.5%	16.2%
Valley of the Shadow	LAB	23.7%	26.3%	13.2%	10.5%	10.5%	5.3%	10.5%	–
	NE*	19.4%	19.4%	13.9%	11.1%	13.9%	13.9%	5.6%	2.8%
Amazon	LAB	–	2.6%	5.3%	15.8%	18.4%	10.5%	23.7%	23.7%
	NE	–	–	–	2.7%	16.2%	35.1%	27.0%	18.9%

*one participant skipped the question in the natural environment.

Table 6. Participants' general evaluation in percentages in response to the question "I consider this website as very good".

Overall, the differences between settings were non-significant. The hypothesis that there is a statistical difference in judgments, here in the participants' general evaluation of digital libraries, between the two settings could therefore not be validated. Put differently, being in a laboratory or in their natural environment does not change a participant's general evaluation of a digital library.

Only one digital library, *Bundesarchiv*, showed an unexpectedly statistically significant difference in the participants' general evaluation of the site between a laboratory and the natural environment with LAB (*M* = 5.26, *SD* = 2.15) and NE (*M* = 6.3, *SD* = 1.37); *t*(63) = -2.45, *p* = .02, two-tailed, *Cohen's d* = .57. The task in the digital library *Bundesarchiv* was the only one that asked for a picture search. Future research should investigate if this was an accident or if the retrieval behavior for pictures is different enough to have an influence on the behavior in different settings.

The *Bundesarchiv* is a digital library for German state history. An additional t-test could not demonstrate any statistically significant difference between participants with a history background and those who did not possess that background. There was also no difference between sex, mood or strong disturbances with regard to the participants' general evaluation of the digital library *Bundesarchiv*. None of these factors was responsible for the difference. There was a slight statistically significant difference for the judgments of this digital library between participants who had additional applications open during the test and those who did not (with LAB (*M* = 5.27, *SD* = 2.2) and NE (*M* = 4.18, *SD* = 2.23; *t*(73) = 2.02; *p* = .05, *Cohen's d* = .47)).

This result is in opposition to the results from the pilot test presupposing that difficult tasks reduce the differences between the settings. Instead, one of the difficult tasks let to a difference in participants' judgments.

Earlier research (Kelly & Gyllstrom, 2011; Tullis et al., 2002) found that remote participants judge the quality of a site more positively. This result could be replicated with the present data set, although it is more a slight tendency instead of a real difference. Table 7 shows that the result of a subtraction of mean values of judgments in a laboratory and a natural environment setting (last column) generally point to one direction (apart from the result in *SSOAR*).

Participants' general evaluation of	Mean (LAB)	Mean(NE)	SD (LAB)	SD (NE)	Sig.	M(LAB)-M(NE)
Perseus	5.37	5.51	1.62	1.71	.71	-0.14
SSOAR	6.42	6.35	1.31	1.44	.83	0.07
Bundesarchiv	5.26	6.3	2.15	1.37	.02*	-1.04
Valley of the Shadow	3.16	3.58	1.99	2.06	.37	-0.42
Amazon	5.95	6.43	1.72	1.07	.15	-0.48

Table 7. Mean values of participant's general evaluations on an 8-point Likert scale (1 = I disagree, 8 = I agree, in response to the question "I consider this website as very good").

Table 7 also shows that in both settings the mean values of the participants' general evaluation of digital libraries are clustered around the center of the scale with a small tendency toward a positive judgment, which might be the result of forcing participants to choose a direction. There was no clear indication that participants strongly approve or disapprove of a particular digital library. Only the digital library *Valley of the Shadow* received obvious negative judgments.

The experiment collected a second judgment after each task that required participants to judge the perceived difficulty of that task (see table 8). The research design expected the first two digital libraries (*Perseus* and *SSOAR*) to be perceived as fairly easy and the next two digital libraries (*Bundesarchiv* and *Valley of the Shadow*) as more difficult, because these digital libraries were more complex and less intuitive in their interfaces and retrieval functions. The last website, *Amazon*, was expected to be perceived at least as easy as the first two digital libraries, because *Amazon* is well known and has very good usability. For all tasks, participants used the non-mobile version of the sites. Table 8 shows that the assumption in the research design matches the participants' perceived difficulty, however also here no significant difference could be detected.

Judgment of perceived difficulty		**I disagree**	**2**	**3**	**4**	**5**	**6**	**7**	**I agree**
Perseus	**LAB**	55.3%	21.1%	10.5%	–	7.9%	–	2.6%	2.2%
	NE	62.2%	24.3%	5.4%	5.4%	–	–	2.7%	–
SSOAR	**LAB**	63.2%	26.3%	7.9%	–	2.6%	–	–	–
	NE	64.9%	18.9%	2.7%	5.4%	–	8.1%	–	–
Bundesarchiv	**LAB**	18.4%	7.9%	21.1%	15.8%	7.9%	10.5%	15.8%	2.6%
	NE	24.3%	16.2%	13.5%	21.6%	8.1%	8.1%	8.1%	–
Valley of the Shadow	**LAB**	–	10.5%	5.3%	7.9%	13.2%	21.1%	15.8%	26.3%
	NE	8.1%	5.4%	18.9%	5.4%	8.1%	10.8%	13.5%	29.7%
Amazon	**LAB**	47.4%	31.6%	7.9%	2.6%	5.3%	–	2.6%	2.6%
	NE	64.9%	10.8%	10.8%	2.7%	2.7%	5.4%	2.7%	–

Table 8. Perceived difficulty per task and setting in percentages in response to the question: "This task was difficult".

Hypothesis 3 states that there is a difference in the participants' judgment—here the perceived difficulty of tasks—between the settings. This hypothesis could not be validated. This means that in order to assess perceived difficulty of tasks, the settings can be equated. This finding contradicts earlier research (for example Adamczyk & Bailey, 2004).

7.4 Difference in the decision-making process

This section examines whether the natural environment setting is a factor that leads to a specific attitude. It tests hypothesis 4, which postulates that distraction in the natural environment makes a difference in the participants' decision-making process. The analysis builds on the Elaboration Likelihood Model by Petty and Cacioppo (1986) which argues that distraction leads to the use of "cues" and less use of "arguments". Distraction makes people use other information. The aim of applying this model is to see in how far the natural environment leads participants to take one route or another.

The research design to test this hypothesis was as follows: after each task, participants were asked how they experienced the search functionality of the digital library (agreement or disagreement on an 8-point Likert scale with the statement that the search functionality was very good) as well as the relevance of the search results (agreement or disagreement with the statement that the search results were very relevant). Afterwards, they judged the professional appearance of the digital library (agreement or disagreement with the statement that the digital library looks professional) and the design (agreement or disagreement with the statement that the design of the digital library is very good). In the Elaboration Likelihood Model logic, functionality and relevance of search results may be characterized as “arguments” (the central route of persuasion), whereas professional appearance and design may be characterized as “cues” (the peripheral route of persuasion). Since the two items “search functionality” and “search results” as well as “design” and “professional appearance” are correlated to a high degree, they were combined to form two sets of single measures: *arguments in task (Perseus, SSOAR, …)* and *cues in task (Perseus, SSOAR, …)*

The variables *arguments in task …* and *cues in task …* for each task were correlated with the participants’ general evaluation or respectively with the perceived difficulty of each task. Using a Fisher transformation, the aggregated variables of each task were then merged to one cumulative variable *arguments* and one cumulated variable *cues* across all digital libraries. The detailed tables with all correlations can be found in the appendix 6. The analysis examines first the judgments of the *perceived difficulty* of tasks and then the judgments on the participants’ *general evaluation* of the digital libraries.

Averaged across all digital libraries and calculated across both settings, a medium correlation between the *perceived difficulty* of tasks and *arguments, r = -.47, p* = .01 emerged. Interestingly, this correlation was less pronounced in the laboratory compared to the natural environment (LAB r = .25, p > .05 vs. NE r = -.49, p = .01 respectively). In contrast, averaged across all digital libraries and calculated across both settings, the correlation between *perceived difficulty* of tasks and *cues* was much lower with r = -.27, p = .03. In this case, the settings did not differ (LAB r = .06, p > .05; NE r = .08, p > .05). Together, these findings suggest that participants in both settings based their judgment of the *difficulty of a task* mostly on *arguments* and less on *cues*.

Averaged across all digital libraries and calculated across both settings, there was a large correlation between the *general evaluation* of sites and *arguments, r = .57, p =* .01. In this case, the settings did not differ (LAB *r* = .54, *p* = .01; NE *r* = .58, *p* = .01). Averaged across all digital libraries and calculated across both settings, however, the correlation between the *general evaluation* of sites and *cues* was even larger (*r* = .79, *p* = .01). Again, the settings did not differ (LAB *r* = .80, *p* = .01; NE *r* = .77, *p* = .01). These results suggest that participants in both settings based their judgments on the *general evaluation* of sites more strongly on peripheral cues such as design than on arguments such as functionality.

Hypothesis 4 postulated that there is a difference in the decision-making process for the participants' judgments between the two settings. Findings suggest that this is not true. There was a difference in the decision-making-process between judgments of the *perceived difficulty* and judgments of the participants' *general evaluation*, but not between the settings. The Elaboration Likelihood Model argues that distraction leads to the use of cues and less use of arguments. This could be demonstrated by comparing participants who had applications open or who have been contacted coincidentally during the test via the internet, for instance, by friends or acquaintances. Averaged across all digital libraries and calculated across both settings, there was a large correlation between the *general evaluation* of sites and *arguments, r = .57, p =* .01. Inspection suggests that this relation was not different between participants that had no applications open during the test (*r* = .58, *p* = .01) and participants that had applications open (*r* = .51, *p* = .03). There was also no obvious difference between participants who have been contacted and those who have not been contacted (with self-reported contact *r* = .53, *p* = .05; without self-reported contact *r* = .58, *p* = .01). These findings may be interpreted as suggesting that distraction—at least as operationalized here—does not change the way participants form general evaluations of a site.

If distraction leads to the use of cues and less use of arguments, then the correlation between the *general evaluation* and *cues* should be different between distracted and non-distracted participants. Averaged across all digital libraries and calculated across all participants, the correlation between the *general evaluation* and *cues* was large with *r* = .79, *p* = .01. There were no obvious differences between distracted and non-distracted participants with participants that had no applications open during the test (*r* = .78, *p* = .01) and participants that had applications open (*r* = .81, *p* = .01); there was also no obvious difference between participants who have not been contacted *(r* = .79, *p* = .01) and participants who have been contacted *(r* = .82, *p* = .01).

These findings indicate that hypothesis 4, which expected a difference between the settings, has to be rejected. It seems that for digital library evaluation, all participants—independent of the setting—rely on cues. It was not the external distraction that made participants focus on the peripheral route—therefore the natural environment does not appear to be a factor that leads to a particular attitude formation process. This finding seems to contradict Petty & Cacioppo (1986) and with that raises doubt about the validity of the model. However, the results primarily show that participants' decision-making process in the laboratory might be more focused on cues than researchers generally expect. The model is not invalid, but the researchers' assumptions about how participants make judgments in a laboratory situation may require adjustment.

There is one danger to the validity of these results: according to the Elaboration Likelihood Model, participants base their decision only on arguments if they are deeply motivated. The fundamental question is if participants in this particular experiment, and more broadly in digital library evaluation studies, were deeply motivated or not. Participants in this experiment were told that the tasks resemble collecting information for a university assessment. In that sense, they were motivated since they could practice how to find information for their university every-day life. Some participants also reported after the experiment that they liked it, because they discovered new sources for their own studies. It is unclear whether that motivation sufficed in the laboratory setting to argue for decision-making based on arguments.

7.5 Difference in the number of page views

Hypothesis number 5 suggests that there is a difference in the number of page views participants needed to complete a task in the laboratory or in the natural environment. The mean values of number of page views spent on tasks were very similar in both settings with an average of 32 page views in the laboratory, with a minimum of 17 page views to complete the tasks and a maximum of 61 page views. In the natural environment, participants needed on average 30.7 page views to complete all tasks with a minimum of 18 page views and a maximum of 59 page views. Table 9 shows the mean values of page views per task and setting.

Page views in…	Perseus	SSOAR	Bundesarchiv	Valley of the Shadow	Amazon
LAB (N = 38)	4.3	3.6	7.4	11.6	5.1
NE (N = 37)	4.4	3.7	6.3	11.4	5.1

Table 9. Mean values of number of page views per task.

Inspection of the mean values suggests that there are no significant differences between the two settings. An independent-samples t-test validates this impression. The test compared the number of page views spent on all tasks (*page views*) for participants in the laboratory and in the natural environment. There was no significant difference in scores for LAB (*M* = 32.05, *SD* = 11) and NE participants (*M* = 30.07 *SD* = 8.14; *t*(73) = .60, *p* = .56, two-tailed). In addition, a multivariate analysis of variance (MANOVA) displayed no significant difference between the number of *page views on the task in …* and the setting. Given a total of 75 participants, an alpha of .05, and an expected medium effect size of .50, the test's power was large enough to detect differences if these were existent (power = .57).

Similar to results by Bowman et al. (2010) and Kirschner & Karpinksi (2010), the statistical tests gave no evidence of a difference in the variable *page views*—measuring the number of page views spent on all tasks—between the settings. Hypothesis number five had to be rejected.

7.6 Difference in the number of successful task completions

Hypothesis number 6 expects that there is a difference in the number of successful task completions between settings. Participants in the natural environment are expected to need more time, but they are not expected to be less successful. Tables 10 and 11 (below) show the task completion results in percentages for the natural environment and the laboratory.

"Completed" means that participants clicked on "task complete" and have found the requested page. "Abandon" means that participants clicked on "abandon task", because they could not find the requested page or had other reasons for abandoning. Finally, "failed" in the context of *Loop11* signifies that participants clicked on "task complete", without having the success URL as last page displayed. The problematic nature of the label "failure" was already discussed in the limitations.

	Perseus	SSOAR	Bundesarchiv	Valley of the Shadow	Amazon
completed	89.5%	92.1%	100%	44.7%	79%
failed	10.5%	7.9%	–	23.7%	18.4%
abandon	–	–	–	31.6%	2.6%

Table 10. Numbers of successful task completions in laboratory in percentages.

	Perseus	SSOAR	Bundesarchiv	Valley of the Shadow	Amazon
completed	97.3%	89.2%	91.9%	48.7%	81.1%
failed	2.7%	2.7%	5.4%	24.3%	18.9%
abandon	–	8.1%	2.7%	27%	–

Table 11. Numbers of successful task completions in the natural environment in percentages.

The "failure" row in table 10 shows that there was a slight tendency in the laboratory to click on "task complete" without having found the requested page (which counted as a "failure"). In the natural environment (table 11) more participants clicked on "task abandon", when they were unable to find the requested page.

The percentages indicate a strong data similarity between the settings. A non-parametric Chi-square test to compare groups was used to examine if there was a difference between the settings. However, for the first three tasks, the data violated the assumptions for a Chi-square test, because zero or near zero values existed. Accordingly, statistical evidence can be gathered for tasks four and five only. A Chi-square test for independence indicated no significant association between the setting and the *number of successful task completions in Valley of the Shadow*, χ^2 (2, n = 75) = .20, p = .91, *Cramer's V* = .05. Similarly, a Chi-square test for independence (excluding the value "abandon task", because of the same reason as above) indicated no significant association between the setting and the *number of successful task completions in Amazon*, χ^2 (1, n = 74) = .00, p = .1, *phi* = .00.

These results suggest that successful task completions were independent of the test setting. Hypothesis 6, which expected a difference in the number of successful task completions between the settings, had to be rejected, in so far as there is no statistical evidence to the contrary. This finding agrees with earlier research (Andreasen et al., 2007; Bowman et al., 2010; Kirschner & Karpinksi, 2010).

At first glance, the high number of "failures" in the *Amazon*-task may be surprising: the task in *Amazon* was intended to be a fairly easy and engaging task at the end of the test—on a website all participants knew. Yet one fifth of all participants failed. In this context, however, it is important to note that a “failure” does not mean that participants did not find the requested information. “Failure” only reflects that the last page was not the requested one. The high number of “failures” is more likely an indication that participants became engaged with the site. The software *Loop11* counts a task as a failure, if the URL before the click on “task complete” is one of the requested success URLs. Participants who found the requested book (= the success URL) and then *continued* to browse before clicking on “task complete”, counted as a “failure”, because the last URL was not the success URL.

An in-depth analysis of the data examines the participants’ “failures” more closely. Appendix 6 shows parts of the clickstream analysis of participants from the *Amazon*-task.[5] The illustration exemplifies the behavior of one participant (highlighted in yellow) at the stage of three previous page views. The task was to find a specific book and to find the number of pages that book had. The participant has found the success URL (first column), but probably overlooks the information on the number of pages. The next page view displays a new search with a change in search terms: the participant added explicitly the search term “page number” to the search terms of the title. Since this did not work out, the participant went back to the original page (the page in column one, now link in column three) and apparently found the requested information at last. Instead of clicking on “task complete”, the participant clicked on product review and then on “task complete”. Since the last URL was the product review and not the requested success URL, this task counted as a “failure”. Without additional interview data, researchers can only speculate about the motives for that behavior.

Further analysis of participants’ behavior revealed that four participants stopped at the search results, having only found the book title, but not the requested page number, which was not accessible at the search results level, but only on the full record page. The participant from appendix 6 looked at product reviews, another one looked at pictures of the book. Seven participants looked for prize and delivery information; they were so intrigued by the book that the task even persuaded them to buy the book. Another participant looked at the imprint. One participant moved even further and started a totally different search in the *Kindle*-Shop. All of these behaviors counted in *Loop11* as “failures”.

[5] Due to size reasons the illustration of the clickstream had to be moved to the appendix 6.

These “failures” show a different user behavior than expected from earlier studies. Researchers expect participants to satisfice and to go for good-enough-solutions. But only two participants in each setting stopped at the search result without having found the requested information. The majority of “failures” in both settings acted in the opposite way: they became engaged by the task and the website, and, after having found the requested information, they continued to browse. There was no difference between the settings in that behavior. One of the aims of digital libraries is to engage users and when users—on a voluntary base—seek additional discoveries on a website, the failure becomes a success. This effect is from now on called *successful failure effect*.

The *successful failure effect* has several implications: first, researchers misinterpret the failures. Second, time becomes an unreliable indicator, because if a participant found the success URL within two page views, the clock continues until the participant clicks on “task complete”. And third, the kind of retrieval tasks researchers developed for laboratory tests might be too engaging in asynchronous remote usability tests in a natural environment, if participants get distracted by the task itself. Future research should examine which stimulation levels lead participants to additional discoveries and therefore to *successful failure effects.*

Loop11 made a choice to use success URLs, but another choice may be to use other forms of measuring task completion such as asking participants to provide the answer to the task explicitly. Then the system could automatically compare answers the participants’ provided with the ones the researcher provided in advance. This approach also has disadvantages, though. For example, the fourth task in the experiment required participants to find the place from which Toni Pastor has written a particular letter. The answer was that he had written the letter from prison. While this task has a clear endpoint, participants might type multiple answers, such as “prison”, “from prison”, “in prison”, which makes that approach no more or perhaps even less reliable than the *Loop11* solution.

In the case of *Loop11*, “failures” emerge because participants behave differently than the software programmers envisioned. Hence, some findings might be interpreted as being an issue of software choice, but this interpretation is too narrow. Asynchronous remote usability tests measure something and need mechanisms for that measurement. It does not matter how software products name the non-completion of a task. What matters is that the way researchers tried to measure task completion does not match users’ behavior in test situations in natural environments.

7.7 Summary

The reported analyses showed that the first two alternative hypotheses (difference in time and variability) could be verified, whereas the alternative hypotheses 3, 4, 5 and 6 (difference in judgments and decision-making, difference in number of page views and number of successful task completions had to be rejected): the findings indicated no difference between the settings. Table 12 below gives an overview of the results. The independent variable for all tests was the setting (values: LAB and NE).

alternative hypothesis	variable(s)	type of test	result	alternative hypothesis was
difference in time (1)	test duration	t-test for independence	significant difference with p < .01	accepted
difference in time (1)	time on tasks	t-test for independence	no difference with p > .05	rejected
difference in time (1)	time on questions	t-test for independence	significant difference with p < .01	accepted
difference in variability (2)	test duration, time on questions	means and extreme values	differences between settings	accepted
difference in participants' judgment (3)	general evaluation	t-test for independence	no difference with p > .05	rejected
difference in participants' judgment (3)	perception of task difficulty	t-test for independence	no difference with p > .05	rejected
difference in the decision-making process (4)	arguments, cues	correlations	no difference	rejected
difference in number of page views (5)	page views	t-test for independence	no difference with p > .05	rejected
difference in number of successful task completions (6)	number of successful task completions in Valley of the Shadow, … in Amazon	Chi-square	no association with p > .05	rejected

Table 12. Summary of findings on differences between the settings.

Hypothesis 1 stated that there is a difference in the time participants needed to complete a test between the settings. There was a statistically significant difference in *test duration* between the two settings, and a difference in *time on questions* between the settings. However, the analysis gave no statistical evidence that the time to complete the tasks (*time on tasks*) was different between the two settings.

Hypothesis 2 stated that there is a difference in the variability of time participants needed to complete a test. The analysis showed that there is indeed a difference in the variability in time between the two settings, especially between slow participants. This difference was caused by the time spent on the questions (*time on questions*).

Hypothesis 3 stated that there is a difference in the participants' judgment of digital libraries. This hypothesis had to be rejected. The analysis showed that there was no difference in judgments between the settings for both the *perception of task difficulty* as well as the participants' *general evaluation* of digital libraries.

Hypothesis 4 stated that there is a difference in the decision-making process for the participants' evaluation. The Elaboration Likelihood Model argues that distraction leads to the use of cues and less use of arguments. Findings suggest that this could not be demonstrated with the present data set. There was a difference in the decision-making process between judgments of *perceived difficulty* and judgments of the *general evaluation*, but not between the settings.

Hypothesis 5 stated that there is a difference in the number of page views participants needed to complete tasks between the settings. This hypothesis had to be rejected. Researches can collect numbers of page views in both settings, because the setting indicated no influence on the page view score.

Hypothesis 6 stated that there is a difference in the number of successful task completions. This hypothesis had to be rejected as well. There was no statistical evidence that the setting is associated with task completion results.

An in-depth analysis of clickstreams disclosed that the definition of "failure" made by *Loop11* has its flaws. If users are intrigued by tasks, then "failures" can become a success. This effect is called *successful failure effect*. These "failures" also showed a different user behavior than known from earlier studies expecting participants to satisfice and to go for good-enough-solutions.

8 Findings and discussion II: control variables

This chapter examines factors of the natural environment and control variables that might help to understand what causes differences between laboratory and natural environment test settings. A more refined understanding of such causes may help researchers to control critical variables.

A key prediction of this experiment was that distraction is responsible for differences between a laboratory and a natural environment setting. It is also possible that other variables causally affected the observed data. In an attempt to assess such influences, several control variables were collected. These can be grouped into the four categories demographics, influences of the environment, internal distraction, and external distractions.

Note that all statistical tests were conducted with three different time variables, namely *test duration*, *time on tasks*, and *time on questions.* In order to improve the ease of reading, emphasis will now be placed on one of these variables, namely *time on questions,* because data between the settings were only different in this variable and not in *time on tasks.* Where illuminating, results on the other two time scores will be reported, too.

8.1 Demographics and experience

Age and sex are standard demographics. There was no statistically significant difference in *time on questions* between females and males (t(73) = -.26, p = .79). The influence of age on time scores was investigated using Pearson product-moment correlations. The relationship between *age* and *time on questions* was of medium size with $r = .40$, $p < .01$ with older participants requiring more time to complete the tasks. A smaller relation exists between *age* and *time on tasks* ($r = .26$).

It is difficult to judge if age causally influenced time scores or whether the participants that were responsible for the difference in time scores between ages were simply the older ones. A follow-up correlation analysis examined the relationship between *age* and *time on questions* for participants that were younger than 27 (average age of participants in both settings was 26). The relationship was minor with $r = .1$, $n = 48$, $p > .05$. In contrast, the relationship between *age* and *time on questions* for participants in both settings who were older than 27 was considerably larger with $r = .51$, $n = 27$, $p < .01$.

Older participants in both settings needed more time to complete the questions. This behavior can have various reasons: younger participants might have been more strongly motivated to complete the questions, because it was one of their first tests. Or younger participants might have been less prone to be distracted—by internal or external distraction—because they felt that they should take the test seriously.

The influence of German or English skills was investigated as well. Participants had to complete two tasks in English digital libraries, first an easy task in *Perseus* and then a more difficult one in *Valley of the Shadow*. Because language skills matter for understanding the sites (*time on tasks*) and for understanding the questions, the variable *test duration* was the reference variable. There was no statistically significant difference in *test duration* between native German speakers and non-natives t(13.51) = -.17, *p* = .12). For the given tasks and participants there was no relation between the four levels of English skills and the *test duration* (*rho* = .08).

Prior experiences with digital libraries might have affected response times, too. These were coded by the number of digital libraries a participant already knew (0–4). A one-way between-groups analysis of variance (ANOVA) with *time on tasks* as a dependent variable showed no statistically significant effect (*F(* 3, 69) = 1.1, *p* = .36).

Prior task-related knowledge was operationalized as the number of semesters, and general education background as the highest degree a participant had acquired. A one-way between-groups analysis of variance with *time on questions* as dependent variable showed no statistically significant difference between the three number of *semester* groups (*F(*2, 72) = .63; *p* = .53). The influence of the *university degree* was investigated using Spearmans correlations and no relationship was found (*rho* = .004, *p* > .97).

Participants studying a specific academic specialization might have had an advantage on the tasks: for instance, participants with an historical background may have had an advantage in the task using the *Bundesarchiv*. Given the nominal nature of these data, no statistical tests could be conducted. However, descriptive inspection of the time scores suggests that time scores did not substantially vary as a function of academic specialization. In conclusion, except for age, no statistical influence by demographic or prior experience variables on *time-of-questions* could be determined. This finding might be surprising, because research in the area of information retrieval repeatedly points to the importance of prior knowledge (see for an overview Vakkari, 1999). The contradictory finding here can be explained by different test approaches: while the information retrieval studies mostly examine tasks that require prior knowledge, this research design explicitly chose test objects and test tasks that did not require specific prior knowledge. Future research needs to assess, which of the two designs correspond better to the standard use of digital libraries: that is how frequently digital library users actually search for information that is so particular that prior knowledge on the topic makes a significant difference.

8.2 Testing environment

There are many factors in a participant's natural environment that might influence the data including the participant's browser or computer type. Both can be logged automatically in an asynchronous remote test, but for reasons of participants' anonymity this information was not collected. In order to limit the questionnaire to a reasonable size, this research collected only information about the nature of the internet connection, the participants' location, and the hour of test completion.

Participants in the natural environment were asked to describe the current place where they completed the test in order to investigate whether a particular location was responsible for differences. Participants were divided into six groups according to their own statements (at home, in a café, in the library, in the library foyer, in a computer-pool in the library and at work). A one-way between-groups analysis of variance with *time on questions* as dependent variable showed no statistically significant difference between the six locations in the natural environment ($F(5,30) = .87$; $p = .51$).

Participants in the natural environment could freely determine the time of participation (assessed on a scale from 0 to 23). The relationship between the *hour of test completion* and the *test duration* was investigated using Pearson product-moment correlations. There was no correlation between the *hour* and the *test duration* for participants in the natural environment, $r = -.03$.

The kind of internet connection when performing the test might influence the time scores as well. Because only five participants used devices classified as "slow" (e.g., mobile or modem connection), these could not be included in formal statistical testing. Descriptive inspection of decision times, however, revealed that participants with a high speed internet connection completed the test faster than participants with a slow internet connection. Excluding the five participants with a slow internet connection, a t-test revealed that there was a statistically significant difference in *time on questions* between participants with a university high speed connection (*M* = 559.29, *SD* = 135.75) and participants with a fast internet connection somewhere else (*M* = 781.4, *SD* = 429.19; *t*(33.6) = -.2.73, *p* = .01, two-tailed). There was also a difference in *test duration* between the two forms of fast internet connection, university network versus high speed internet somewhere else, *t*(39.5) = -2.18, *p* = .04. A follow-up t-test, however, revealed that there was no statistically significant difference in *time on tasks* between the fast DSL users and the university network users.

The natural environment was defined to be everywhere else except in the laboratory. Participants could choose where they wanted to take part in the test. The results of the tests above indicate that *time on questions* differs when participants did not use the university high speed internet connection. It is interesting that a difference in time only occurs during the questionnaire and not on the tasks. One could argue that as long as participants are occupied with tasks, the kind of internet connection does not matter. As soon as participants are less heavily engaged, e.g. during the questionnaire, participants at home seem to be more likely distracted. In consequence, the influence of the home environment is bigger than the influence of public university spaces.

8.3 Internal distractions

Participants can be distracted by their environment, but also by their own state of mind. This research collected information on three internal distractions—mood, daydreams, and level of attention. All three variables were assessed as self-reports.

The relationships between the three internal distractions and time on questions were investigated using Pearsons product-moment correlations. There was no relationship between the different moods (good or bad temper on an 8-point Likert scale) and the *time on questions* (r = .01, p > .05), the degree of *daydreams* (r = .02, p > .05) or the *level of attention* (r = -.11, p > .05).

The variable *time on questions* was not related to internal distraction. It seems that collecting information about internal distractions would make sense for outlier analyses in order to detect participants who were in a really bad mood and did not concentrate at all on the test, but it is unlikely that these participants will honestly supply the information. In that sense, self-reported information about internal distraction is of lower value than other information described in this research.

8.4 External distractions

Chapter six reported that a number of technical problems occurred during the test. A technical problem could be something serious like a server outage or something more unproblematic like minimally longer loading times. An independent-samples t-test was conducted to compare the *time on questions* for participants who had problems and who did not. There was no significant difference on *time on questions* between the two groups (*t*(35) = -.73, *p* = .47).

The central research hypothesis that guided the present research postulates that the difference in *test duration* between settings is due to distraction. As one way to operationalize distraction, the number of additional open programs was assessed. In the analyzed sample, 64.9% of participants in the natural environment had another program open during the test. T-tests revealed no significant difference on any of the time scores: for *test duration*: *t*(35) = .14, *p* = .89; for *time on tasks*: *t*(35) = .55, *p* = .59; for *time on questions*: *t*(35) = .68, *p* = .50.

An additional t-test revealed no significant difference in *time on questions* with participants who indicated that they had ignored the open programs and participants who looked at the programs (*t*(34) = .33, *p* = .75). In addition, descriptive inspection suggests that there was not one particular external program (like *Facebook* or an Email client) that caused differences in *time on questions*.

Based on these results, open programs or even an active switch between the online test window and other programs likely do not alter time scores. With the necessary note of caution, one may conclude from these findings that closing additional programs before a test—as many researchers require—may not be necessary. In fact, there may even be a risk that participants get annoyed by the requirement and therefore start the test with a rather negative mood.

In addition to open programs, the natural environment enables external distractions such as contacts by other individuals. In the experiment, participants were asked whether they had been contacted during the test, either by phone, SMS, or face-to-face. The following tests investigated whether contacts resulted in no differences on all time scores as well.

With respect to *time on questions*, there was a non-significant tendency, suggesting that contacted participants (*M* = 1088.9, *SD* = 602.93) took longer than undisturbed ones (*M* = 674.93, *SD* = 195.94; *t*(9.71) = 2.13, *p* = .06, two-tailed). The influence on *time on questions* by the frequency of contacts was investigated using Pearsons product-moment as well. There was a large positive correlation between the two variables, *r* = .52, *n* = 37, p < .001. This means that a higher contact frequency can be associated with slower completion rates on the questionnaire.

A second independent-samples t-test was conducted to compare the *test duration* for participants who had been contacted and the ones who were undisturbed. There was a significant difference in the *test duration* for undisturbed participants (*M* = 1297.11, *SD* = 458.3) and participants who had been contacted (*M* = 2017.20, *SD* = 773.86; *t*(35) = 3.5, *p* < .001, two-tailed).

A third independent-samples t-test was conducted to compare the same groups on *time on tasks*. There was a significant difference in *time on tasks* for undisturbed participants (*M* = 622.19, *SD* = 301.53) and participants who had been contacted (*M* = 928.30, *SD* = 276.89; *t*(35) = 2.80, *p* = .01, two-tailed).

The variable *time on tasks* which is crucial for the measurement of the time-on-task-performance was not significantly different between the laboratory and the natural environment and the conclusion was that for time-on-task-performance measurements the setting does not matter. The last results modify this statement, because there was a statistically significant difference between disturbed and undisturbed participants in *time on tasks*.

One limitation of this result is the small sample of participants who were contacted during the test (10 participants out of 37). This requires further research in which contacts might be stimulated or a larger sample from an asynchronous online test in which enough contacts occur.

The strong influence of contacts causes a real problem for online test designs: researchers can ask participants to close programs before the test starts. They even might ask them not to take phone calls (knowing that a participant is unlikely to ignore a noisy phone call). But researchers will never be able to control face-to-face contacts, for example a mother taking care of her crying children or someone coming into a room asking for advice: the strongest influence on online data in a natural environment is caused by something that researchers are unable to control. Researchers can only try to collect as much information on contacts as possible to be able to interpret the data.

The central research hypothesis that distraction is at the origin of the difference between the settings could be demonstrated. Multitasking, which seemed the obvious distraction in the natural environment and therefore a risk to online tests, does not alter the data in a significant way. Contacts, on the other hand, change the data in a significant way. Hence it does not matter where we test, but it matters what happens during the test.

8.5 Summary

This chapter examined influences on data which might be the cause for the differences between a laboratory and a natural environment test setting. Multitasking does not alter the data in a significant way. Contacts, on the other side, change the data in a significant way.

Based on findings from several disciplines, a catalog of possible effects was tested in order to provide guidance on potentially influencing factors in a natural environment setting and on those that can be neglected in future studies. A number of control variables could be found that demonstrated no influence on the data:

- *Gender* did not alter time scores.
- *Language skills* did not alter time scores.
- *Prior experiences with the digital libraries* did not alter the *time on tasks*.
- Prior task related knowledge (operationalized by *number of semesters*) or the education background (operationalized by *university degree*) did not alter the *time on questions*.
- Internal distractions (operationalized by *mood, level of attention* and degree of *daydreams*) were not related to the *time on questions*.
- The existence of additional open programs or even an active switch between the online test window and other programs did not alter time scores.
- *Contacts* did not alter the *time on questions*.

Significant influences on the time scores were:

- *Age* had a medium influence on time scores. Older participants in both settings needed more time to complete the questions.
- *Time on tasks* and *test duration* was different between the *kind of high speed internet connection* outside the university network and within the university internet network.
- *Contacts* alter time scores. There was a statistically significant difference in *test duration* and in *time on tasks* between contacted and undisturbed participants.

Several results were significant either in *time on tasks* or in *time on questions*, but rarely in both. For example contacts during the questions did not significantly alter the data; contacts during the tasks had a huge impact. The overall variable *test duration* masks events that happen either during the tasks or during the questions. This means that *test duration* is a fuzzy variable for user behavior tests and needs to be considered carefully for its individual components' impact on test results.

Table 13 gives an overview of the findings:

	Independent variable	**Dependent variable(s)**	**Test**	**Result**
Demographics and Experience	age	time on tasks, time on questions	Pearson	medium correlation
	age > 27	time on questions	Pearson	large correlation
	experience with the digital libraries	time on tasks	ANOVA	not significant
	gender	time on questions	t-test	not significant
	university degree	time on questions	Spearman	no correlation
	level of English skills	test duration	Spearman	no correlation
	native German speaker and non-native	test duration	t-test	not significant
	number of semesters	time on questions	ANOVA	not significant
Environment	location in the NE	time on questions	ANOVA	not significant
	hour of test completion	test duration	Pearson	no correlation
	university internet connection and other fast speed connection	test duration, time on questions	t-test	significant
	university internet connection and other fast speed connection	time on tasks	t-test	not significant
Internal distraction	level of attention	time on questions	Pearson	no correlation
	daydreams	time on questions	Pearson	no correlation
	mood	time on questions	Pearson	no correlation

	Independent variable	Dependent variable(s)	Test	Result
External distraction	contacts	time on questions	t-test	not significant
	contacts	test duration, time on tasks	t-test	significant
	frequency of contacts	time on questions	Pearson	large correlation
	technical problems	time on questions	Pearson	not significant
	"looked at programs"	time on questions	t-test	not significant
	particular program	time on questions	t-test	not significant
	open programs	test duration, time on tasks, time on questions	t-test	not significant

Table 13. Summary of findings on control variables.

9 Conceptual framework for online user studies in natural environments

In asynchronous remote usability tests, researchers and participants are separated in space and time. During the tests, researchers are unaware of disruptive events on the participants' side outside the test browser window—unless they collect that information. This information can be collected in the form of variables which help to describe a participant's natural environment during a test situation.

The research in this dissertation collected several variables that might have a major influence on completion time scores and are therefore crucial for valid data interpretation. While collecting this large amount of variables made sense for this experiment, it might not necessarily be the case for all online user studies in natural environments: a few well selected variables might be sufficient to collect all the data that is needed. This experiment included a section where the participants performed tasks (about 10 minutes) and an equally long section on questions. Usually, online studies contain more or longer tasks. In consequence, the question part of the test must be shorter in order to avoid long experiments, which would increase the participants' dropout rate.

This chapter organizes the experimental results into a framework for the kind of variables that need to be collected in online user tests in a natural environment to be able to interpret the data from these tests. The aim of the framework is to offer researchers a mechanism to determine which variables matter for data collection. Three types of variables can be distinguished:

- *Core* variables

  These variables are indispensible for statistical tests and ought to be collected in online studies in a natural environment.

- *Informative* variables

  These variables do not always influence the statistical results, but can be important if researchers want to examine an individual users' behavior.

- *Additional* variables

  These variables have not shown any statistically significant influence on the data within this experiment, but depending on the search task or sample, they could be useful for data analysis.

9.1 Measurement of time

Completion time has been an important indicator to collect in online user studies in digital libraries. Despite this common practice, time scores are risky as a measurement of performance, because the interpretation of that performance is fuzzy. A result such as "300 seconds on a task" can have too many meanings to be used for a clear performance measurement (see chapters 5, 7 and 8). Based on the findings in this work, it is safest to collect time as an indicator for "events" that might have influenced users' behavior.

Time on tasks (the time participants spent on all tasks) as well as the *time spent on the task in …* are core variables. If strong disturbances like contacts occur during a test situation, the variable *time on tasks* is affected and can indicate that something happened. These indications can be seen at the extremes: very slow participants could indicate an external disturbance like a contact and extremely fast participants could indicate a technical problem like a server outage (which made them give up). In that sense, the variable offers a reasonable way to measure whether a disturbing event took place during the tasks. The advantage of the variable *time on tasks* is that most software packages provide it and that it does not need to be asked in the survey section. It is also defined as a core variable, because it was not affected by the natural environment itself. The statistical tests showed no difference in *time on tasks* between a laboratory and a natural environment and the power test indicated that the data was the same between the two settings.

The variables *time on the task in …* are core variables, because they help researchers to detect problematic tasks or events. The outliers' description showed that participants were generally tardy on only one part of the test, rather than across all parts of the test, suggesting either an interruption during a particular task or interaction problems with particular websites.

Time on questions measures the time that participants spent completing the questionnaire. The variable is categorized as an *informative* variable, because it can indicate whether participants were distracted during the questionnaire. Data from this variable were significantly different between the two settings, with participants in the natural environment usually taking longer to complete the questionnaire. The variable is not a core variable, because completion times on the questionnaire can indicate a distraction, but they appear not to alter the answers to the questionnaire. It matters little if participants need more time, if the results are still the same.

Test duration examines the time participants need to complete the whole user test. The variable can be collected as an additional variable to examine if researchers expectations of test time was correct. The findings suggest that *test duration* is a fuzzy variable which is determined by *time on tasks* and *time on questions*. It consequently is classified as additional variable.

9.2 Measurement of task completion

Page views on tasks is a core variable for online user studies. There was no statistically significant difference between the settings in the number of page views, which means that the variable was likely not affected by the natural environment itself. The number can be used as a valid way to assess the difficulty of task completion (note that *page views* are collected automatically and should be distinguished from the variables on judgments). In general, more page views are equivalent to a more difficult task, and the more page views users needed, the more they struggled with a task. If individual participants show an uncommonly large number of page views, the number of page views can indicate a *successful failure effect* (chapter 7).

Individual page views in the form of log data allow researchers to reconstruct a session and therefore are classified as core variables as well. The overall number of page views only indicates extreme cases, while the log data of individual page views allow an in-depth analysis of users' behavior.

Task-answers, i.e. participants' written answers to the tasks, is an informative variable. This work did not require participants to provide the answer and therefore might have permitted participants to abandon the task more easily or to click on "task complete" without the required answer. Task completion can be measured by comparing the provided answer with the required one. This can only be done for tasks that have a clear answer like a number or a name, which a system can automatically check. The variable *task-answers* is categorized as an informative variable, although this way of measuring task completion has flaws, too. For more complex tasks, an automatic measurement of successful task completions is currently not possible using this kind of tests.

Task completion as measured by *success URL* is an additional variable. It can be useful for a quick assessment of success rates. There is, however, a potential risk of reaching wrong conclusions when considering this variable. The automatic measurement expects participants to click on "task complete" when participants have completed the task and not before or afterwards. The experiment showed that some participants clicked on “task complete” before the required URL was displayed (potentially an indication of users’ satisficing). Their task completion was marked as a “failure”. Without additional information, this behavior is hard to interpret. Some participants also found the requested URL and continued to browse and to engage with the sites (*successful failure effects)*. Their behavior was also marked as a “failure”. Until these phenomena are taken into consideration, using the *success URL* as a variable for task completion can only have an additional value.

9.3 Measurement of judgments

The two variables (*search functionality* and *relevance of search results*) are informative variables. Together with the variable *perceived difficulty* of tasks, they collect judgments. The analysis showed that the three variables are strongly correlated and therefore it suffices to collect only one of the three. Since not all participants necessarily use the search functionality to solve the task, *perceived task difficulty* is defined as the core variable in this group.

A general evaluation of a website, for example a rating, is an informative variable. The variable was not affected by the settings and it can provide useful information for researchers aiming at collecting judgments. There is a risk that the variable is strongly influenced by what Petty and Cacioppo (1986) called “cues”, which means that the judgments only measure the impression participants have of the design of the website including the website’s professional appearance.

Judgments about a participant’s own distraction during the test are additional variables, because participants tended to misjudge their own level of distraction. This variable is useful to compare the perceived level of distraction with actual distraction, but it provides too little information about real distraction levels to justify a collection in every online user study. The variables *daydreams*, *mood* and *level of attention* are classified in this category.

9.4 Measurement of control variables

Information about *contacts* is a core variable. The findings have shown that contacts alter the variable *time on tasks* and that some knowledge about the existence of contacts is essential for data interpretation. The *frequency of contacts* is also a core variable. Future test designs might also consider break buttons, so that participants could indicate at what time a contact occurred. Both whether this convention would be adopted by the participants as well as the efficiency of the break button approach need to be validated.

Technical problems are a core variable as well, even if there was no statistically significant difference in participants with and without problems between the settings. In an asynchronous remote usability tests, researchers receive no information about what happened on either the server or on the participants' side during a test, but the outlier description suggests that knowledge about technical problems can be crucial for an individual participant's behavior. An example of a situation in which the knowledge about technical problems can be essential for data interpretation is a single page view, where many seconds were spent on that task.

As important as knowing about the occurrence of a problem is having a description of what occurred. Participants in this study reported all kinds of problems during the test, some of which were genuine technical problems (server outage or problems with search functionality), some of which were usability issues (long loading times or bad design). Knowledge about the kind of problems is a core variable, because it helps researchers to distinguish between actual technical problems and usability issues.

Language skills are core variables as well. Participants needed more time to complete the tasks and the questionnaire if their native language was not the test language. This is an important factor for international tests. A misleading interpretation might suggest that participants from specific countries need more time to search, when in fact they only needed more time to understand the tasks and the questions. There was no statistically significant difference between the settings where participants were native speakers or where participants indicated high or low English facility. Language does not matter from a statistical point of view (at least within this experiment) but it matters for interpreting individual user's behavior as shown in the outlier description.

The kind of *internet connection* is a core variable as well, because the connection had a considerable influence on the data. Completion time scores might be misleading if the reason for a faster test completion cannot be attributed to participants' skills, but on their kind of internet connection.

Age was an influential variable in this experiment and is therefore a core variable for data collection. The variable might be less important if all participants were within a very small range or were generally younger. The strong influence of age on time scores mattered for older participants (>27) versus younger ones.

Information about *open programs* during the test is an informative variable. The findings gave no evidence that the fact of having a program open altered the data between the settings. The same was true for particular types of programs or the number of open programs. Information about this kind of multitasking can be collected, but is not crucial for data interpretation. The outlier description has shown that knowledge about open programs allows researchers to speculate about what happened during the test, but the influence of other programs on the time scores was less obvious than, for example, technical problems or contacts.

Other demographic and experience indicators such as the *education background, the number of semesters* at the university, *digital library knowledge, gender* or *academic specialization* must be grouped as additional variables, because they demonstrated no influence on the participants' behavior on the time scores.

Finally, *location of test completion* and *time of test completion* can be grouped under additional variables as well, because they did not lead to different results on the time scores. Both might be of great interest in future studies that take a deeper look into the participants' natural environment and try to find out how, where and when participants use a digital library. The difference in *time on questions* between the kinds of internet connection suggested that the influence of the home environment might be bigger than the influence of public university spaces. If researchers continue to explore the natural environment in which the users' interaction with digital libraries takes place, and in which these users participate in tests, variables about the natural environment, such as details about the location and the time of test completion, could become a major focus of research—as were the variables on distraction in this work.

Figure 12 shows the three types of variables with the core variables at the center of the framework (yellow circle), the informative variables as a second circle (red) and the additional variables as an outer circle (blue). Variables in the quadrants I and III can be produced by software in asynchronous remote usability tests, while variables in the quadrants II and IV require survey techniques. The experiment focused on quadrants I (time scores) and II (control variables). The list of variables in quadrants III (task completion) and IV (judgments), as well as possible further control variables that might influence task completion and judgments is likely not exhaustive at this stage. The framework is a first attempt to categorize variables for data collection in online user studies in digital libraries.

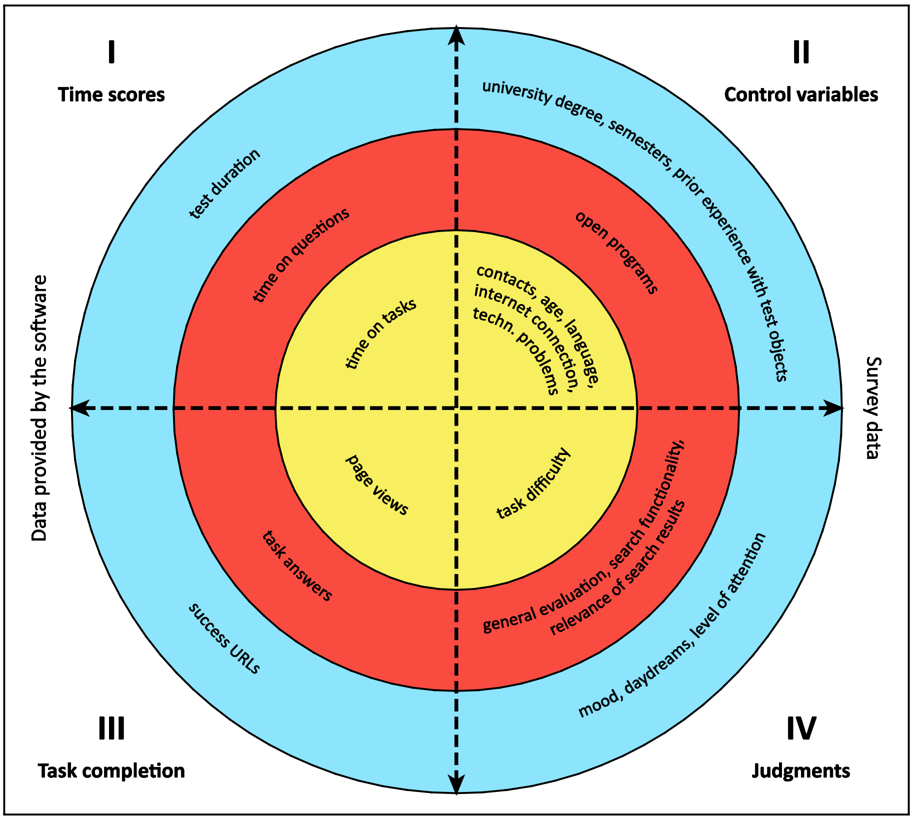

I
Time scores
II
Control variables
III
Task completion
IV
Judgments
Data provided by the software
Survey data
test duration
time on questions
time on tasks
university degree, semesters, prior experience with test objects
open programs
contacts, age, language, internet connection, techn. problems
page views
task answers
success URLs
task difficulty
general evaluation, search functionality, relevance of search results
mood, daydreams, level of attention

Figure 12. Conceptual framework for online user studies in natural environments: type of variables with *core* variables in yellow, *informative* variables in red and *additional* variables in blue.

10 Conclusion

User studies in digital libraries face two fundamental challenges. The first is the necessity of running more user studies in an online environment. Online studies enable researchers to be separated from their participants in space (synchronous tests) and/or in time (asynchronous tests). This need for more online studies is coupled with a second need, a demand to test under realistic conditions outside of laboratories in users' natural environment.

Asynchronous remote usability tests are a methodological approach that might answer both needs: they allow participants to take part in a study at a time and place of their choice, often in the participants' natural environment. Any chosen place, however, might be noisy. Distractions are ubiquitous in a user's natural environment. An awareness of the potential influences of these distractions on users' behavior during test situations is of great importance, because the validity of a study depends on the quality of the data. If an instrument allows systematic mistakes in measurements because of distractions, the validity is at risk. This work examined if distraction in the users' natural environment produces a systematic mistake in digital library studies that take place at a time and location of participants' choice. As the answer is yes for certain variables, results and their interpretations need to be considered carefully when based on these conditions.

In order to investigate the existence of distractions during online user studies in digital libraries and to analyze the influence(s) of that distraction, a psychological experiment was set up. It examined completion time scores between randomly assigned participants in a laboratory and participants in their natural environment. Both groups completed the same asynchronous remote usability test, which consisted of five retrieval tasks in four digital libraries and in an online-shop serving as control site. The participants were unaware that two settings existed and they did not know that the user study was about distraction. Survey data on the participants' distraction level during the test were collected. Based on earlier research, it was expected that participants in the natural environment would need more time to complete the same test, because distraction leads to an increase in completion time. The null hypothesis for this work stated that the data on the time scores between the two settings were the same.

The results of the experiment showed that participants were highly distracted. Most participants in the natural environment had other programs open during the test situation and many were contacted during the test. The analysis showed that this distraction affected the time spent on the test and that participants in their natural environment needed more time to complete the same test. In addition, there was a large variability in participants' time scores in the natural environment. The setting did not affect either successful task completions or the participants’ judgments. There was a statistically significant difference in the time spent on the questions between participants in a laboratory and participants in a natural environment, but there was no statistically significant difference between the settings in the time participants needed to complete the tasks themselves. This means that as long as participants were occupied with tasks, the setting appeared to matter less. Multitasking as the core origin for the distractions in the natural environment caused no statistical difference in completion time scores between the settings. The same was true for technical problems such as server outages.

The experiment suggested that data were often not very different between the two settings. The laboratory and natural environment settings were actually more alike than might have been expected. The findings on the decision-making process of the participants’ judgments also illustrated this: there was a difference in the decision-making process between judgments of the perceived difficulty and judgments of the participants’ general evaluation, but not between the settings.

For those participants who faced a strong disturbance, however, for example someone calling them, the situation changed. There was no evidence for a difference in the variable *time on tasks* between a laboratory and a natural environment, but participants who were contacted during the test spent a statistically different amount of time on the tasks than those participants who were not contacted. This result makes asynchronous remote usability tests in a natural environment difficult, because even if researchers were able to dissuade participants from multitasking, controlling external contacts is very hard.

When researchers cannot control a situation (or choose to refrain from controlling it, because they want to keep the setting realistic), they need to have information that lets them retrace the situation as much as possible. Researchers need to know what happened on a participant’s side during the test to be able to give meaning to the data. Without a description of the setting, researchers risk interpreting the data incorrectly or superficially. For example, a long completion time on one task could mean that a participant struggled with that particular task or that this participant was contacted at the beginning of the task, paused, and afterwards finished the task quickly. Information about what happened in the natural environment is indispensable in order to interpret the data about time scores in asynchronous remote usability tests.

The conceptual framework for online user studies in natural environments, which was developed based on the findings of the experiment, suggests three types of variables that need to be collected: *core* variables that are necessary for data collection, *informative* variables that can help to interpret individual users' behavior, and *additional* variables that are not required but still can be useful for particular research questions. One avenue for future research would be the assessment of circumstances under which additional variables become *core* variables.

The framework further suggests a matrix of four thematic orientations. These orientations are variables on the measurement of task completion, the collection of judgments, the measurement of time as well as the collection of control variables. Control variables go far beyond purely demographic data. While the experiment in this dissertation focused on the time and control variables, some issues regarding the measurement of task completion arose as well.

In the experiment, some participants showed a behavior labeled a *successful failure effect*. It means that participants got engaged by the task and continued to browse after having completed the original task. The software *Loop11* evaluates a successful task completion by comparing a participant's last page view on a task with a predefined success URL. If a *successful failure effect* occurs, this automatic form of measuring task completion fails. In that sense, some numbers are simply a software problem. *Loop11* also defines the variable *time on tasks* as the time that was spent on a task until the participant clicked on "task complete" or alternatively on "task abandon". In consequence, the *successful failure effect* leads to unreliable scores on the time spent on the tasks, because participants' browsing leads to an increase in time. Again, this is a software problem. Other data collection or interpretation approaches are not as easily solved. The technological side can be changed easily, but the methodological aspect needs to be discussed within the research community and should not be left to software engineers. The essential question is how effectively researchers in this field can adapt the method to their needs.

One part of this discussion might be the degree to which researchers could dispense with time scores on the tasks, since they seem to be an unreliable indicator of task performance in asynchronous remote usability tests. Conversely, one might discuss how scores on the variable *time on tasks* could be collected to be more meaningful. A second part of this discussion could be whether software needs to be required to collect variables on task completion in an automatic way, because finding an appropriate way to measure successful task completions automatically is difficult. Researchers might be forced to decide to dispense with automatic measurements entirely and to accept the fact that testing large samples is not possible without devoting the time and money needed to examine the results one by one. Researchers need to discuss to what degree—if at all—more open-ended information tasks can be measured. If researchers decide on tasks with one clear endpoint, they must realize that these tasks risk becoming artificial—which leads away from a realistic test setting that allows insight into users' real behavior.

Researchers need knowledge about the influence the test setting has on the test results. The experiment has shown that there are variables that are not presently collected in user studies on digital libraries that are essential for data interpretation. The outlier description has shown that knowledge about server outages can be essential for the interpretation of individual participants' behavior. The strong influence of contacts on users' time scores demonstrates the need to collect information about contacts in future user studies. Without input from these variables, there can be no valid interpretation of results.

Bustamante (2010) stated that *Loop11* offers objective and clear data. This is only partly true. One could argue that the completion time scores and the page views provided by *Loop11* are objective, because they objectively report the pure numbers, for example that participant 61 needed 332 seconds on task two. The interpretation of that output, which is partly offered by *Loop11*, causes the problem. The data are objective, but the interpretation is not always clear. Knowledge about distraction is a major key for interpreting data from a natural environment test setting, but other factors are also important. The kind of internet connection a participant used, for example, had a large impact on the time spent on the questionnaire. The natural environment likely holds many more variables that exert a secret influence and these variables need to be collected in order to allow an adequate interpretation of the data.

This dissertation focused on distraction and its potential implications for user studies in digital libraries. The research used digital libraries as the object of study in order to highlight their need to face distraction as a topic of research, but the results are probably transferable to other objects as well. In this study, the data between a laboratory and a natural environment differed only on a few scores. This dissertation demonstrated that it may not matter where we tests, but it certainly matters what happens during the test. It matters to know what might have affected the data. Even if the data looks the same, the meaning can be very different. Good online user studies in natural environments are more than assigning tasks and collecting participants' judgments. A core part of these tests should involve collecting information about the test environment. The danger of data collection in a natural environment is not that events might occur, but that researchers know nothing about them.

11 References

11.1 Bibliography

Adamczyk, P. D. & Bailey, B. P. (2004). If not now, when?: The effects of interruption at different moments within task execution. In: *Proceedings of the CHI '04 conference* (pp.271–278). New York, NY: ACM.

Agosto, D. E. (2002). Bounded rationality and satisficing in young people's Web-based decision making. *Journal of the American Society for Information Science and Technology*, *53*(1), 16-27.

Akselbo, J. L., Arnfred, L., Barfort, S., Bay, G., Bagger Christiansen, T., Hofman & Hansen, J., et al. (2006). *The hybrid library: from the users' perspective: A report for the DEFF project "The loaners' expectations and demands for the hybrid library".* Retrieved from http://www.statsbiblioteket.dk/summa/fieldstudies.pdf

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O. & Stage, J. (2007). What happened to remote usability testing?: an empirical study of three methods. In: *Proceedings of the CHI '07 conference. San Jose, California, USA* (pp.1405–1414). New York, NY: ACM.

Andrzejczak, C. & Liu, D. (2010). The effect of testing location on usability testing performance, participant stress levels, and subjective testing experience. *Journal of Systems and Software*, *83*(7), 1258-1266.

Arms, W. Y. (2000). *Digital libraries*. Cambridge, MA: MIT Press.

Baker, R. M. S., Kiris, E. & Vasnaik, O. (2007). Testing remote users: an innovative technology. In: N. Aykin (Ed.). Usability and Internationalization. Proceedings of the HCII '07 conference. Lecture Notes in Computer Science: Vol. 4559. (pp.235–242). Berlin: Springer.

Baltar, F. & Brunet, I. (2012). Social research 2.0: virtual snowball sampling method using Facebook. *Internet Research*, *22*(1), 57–74.

Bargas-Avila, J. A. & Hornbaek, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In: *Proceedings of the CHI '11 conference* (pp.2689-2698). New York, NY: ACM.

Bartek, V. & Cheatham D. (2003). *Experience remote usability testing, part 1: Examine study results on the benefits and downside of remote usability testing*. IBM DeveloperWorks. Retrieved from http://www.ibm.com/developerworks/web/library/wa-rmusts1/

Bates, M. J. (2010). Information Behavior. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of Library and Information Sciences* (Vol. 33rd ed., pp. 2381–2391). New York: CRC Press.

Batra, S. & Bishu, R. (2007). Web usability and evaluation: issues and concerns. In: N. Aykin (Ed.). Usability and Internationalization. Proceedings of the HCII 2007 conference. Lecture Notes in Computer Science: Vol. 4559. (pp.243–249). Berlin, Heidelberg: Springer-Verlag.

Bayram, Ö. & Doğan, A. (2006). An evaluation of faculty use of the digital library at Ankara University, Turkey. *The Journal of Academic Librarianship*, *32*(1), 86–93.

Beschnitt, M. (2009). *Standardisierte Fragebögen zur Messung der Usability/User Experience – derzeit wieder in Mode? | usabilityblog*. Retrieved from http://www.usabilityblog.de/2009/09/ standardisierte-fragebogen-zur-messung-der-usabilityuser-experience-derzeit-wieder-in-mode/

BIX Handbuch – WB 2011 (2011): Erhebungsunterlage für den BIX für wissenschaftliche Bibliotheken 2011 (Berichtsjahr 2010). Retrieved from http://www.bix-bibliotheksindex.de/

Bogros, O. (2003). La bibliotheque electronique de Lisieux-Etat des lieux: mythes et réalités. *Bulletin des Bibliothèques de France*, *48*(4), 45–48. Retrieved from http://bbf.enssib.fr/consulter/bbf-2003-04-0045-008

Boldt|Peters. (2005). *Remote testing versus lab testing*. Retrieved from http://www.boltpeters.com/articles/versus.html

Bolton, R. J. & David, J. H. (2002). Statistical fraud detection: A review. *Statistical Science*, *17*, 2002.

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *53*, 225–250.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *56*(1), 71–90.

Bowman, L. L., Levine, L. E., Waite, B. M. & Gendron, M. (2010). Can students really multitask? An experimental study of instant messaging while reading. *Computers & Education*, *54*(4), 927–931.

Brasel, S. A. & Gips, J. (2011). Media multitasking behavior: Concurrent television and computer usage. *Cyberpsychology, Behavior, and Social Networking*, *14*(9), 527–534.

Brewer, M. B. (2000). Research design and issues of validity. In: *Handbook of Research Methods in Social Psychology* (pp.3–16). Cambridge University Press.

Brush, A. B., Ames, M. & Davis, J. (2004). A comparison of synchronous remote and local usability studies for an expert interface. In: *Proceedings of the CHI '04 conference* (pp.1179-1182). New York, NY: ACM.

Bruun, A., Gull, P., Hofmeister, L. & Stage, J. (2009). Let your users do the testing: A comparison of three remote asynchronous usability testing methods. In: *Proceedings of the CHI '09 conference* (pp.1619–1628). New York, NY: ACM.

Buchanan, E. A. (2004). *Readings in virtual research ethics: Issues and controversies*. Hershey, PA.: Information Science Publication.

Bullard, J. & O'Brien, H. L. (2011). Online synchronous interviewing of the info-savvy. In: *Proceedings of the iconference '11* (pp.649-650). New York, NY: ACM.

Bustamante, J. (2010). La herramienta de tests de usabilidad a distancia Loop11. *Profesional de la Informacion*, *19*(4), 425–430.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, *31*(2), 191–213.

Casden, J. (2011). *Suma: A Mobile Space Assessment Toolkit*. Retrieved from http://www.lib.ncsu.edu/dli/projects/spaceassesstool/

Case, D. O., Andrews, J. E., Johnson, J. D., & Allard, S. L. (2005). Avoiding versus seeking: the relationship of information seeking to avoidance, blunting, coping, dissonance, and related concepts. *Journal of the Medical Library Association*, *93*(3), 353–362.

Chase, L. & Alvarez, J. (2000). Internet research: The role of the focus group. *Library & Information Science Research*, *22*(4), 357–369.

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. In: *Proceedings of the IUI '01 conference* (pp. 33–40). New York, NY: ACM.

Clemmensen, T. & Plocher, T. (2007).The cultural usability project: studies of cultural models in psychological usability evaluation methods. In: N. Aykin (Ed.). Usability and Internationalization. proceedings of the HCII 2007 conference. Lecture Notes in Computer Science: Vol. 4559.

Cohen, J. W. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, M. J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N. J., & Zhang, X. (2011). Task and user effects on reading patterns in information search. *Interacting with Computers*, *23*(4), 346–362.

Czerwinski, M., Cutrell, E. & Horvitz, E. (2000). Instant messaging and interruption: influence of task type on performance. In: *Proceedings of the Australian Computer-Human Interaction conference – OZCHI* (pp.356–361). Coffs Harbour, Australia, Southern Cross University.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). *Applied Social Research Methods Series: Vol. 26*. Thousand Oaks, Calif: Sage Publications.

Diekmann, A. (2005). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (13th ed.). Reinbek bei Hamburg: Rowohlt-Taschenbuch-Verlag.

Dixon, B. E. (2009). Enhancing the informatics evaluation toolkit with remote usability testing. In: *Proceedings of the AMIA Annual Symposium.*

Fagan, J. C. (2010). Usability studies of faceted browsing: A literature review. *Information Technology and Libraries*, *June*, 58–66.

Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: A flexible statistical power analysis program for the tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In: C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology* (pp.74–97). Thousand Oaks, CA, US: Sage Publications.

Ferran, N., Casadesus, J., Krakowska, M. & Minguillon, J. (2007). Enriching e-learning metadata through digital library usage analysis. *Electronic Library*, *25*(2), 148–165.

Frey, D. & Stahlberg, D. (1993). Das Elaboration-Likelihood-Modell von Petty und Cacioppo. In: D. Dauenheimer, D. Frey & M. Irle (Eds.), *Kognitive Theorien. Theorien der Sozialpsychologie* (pp.327–360). Bern: Huber.

Fried, C. B. (2008). In-class laptop use and its effects on student learning. *Computers & Education*, *50*(3), 906-914.

Gardner, J. (2007). Remote web site usability testing: benefits over traditional methods. *International Journal of Public Information Systems*, *2*, 63–72.

Geertz, C. & Luchesi, B. (2009). *Dichte Beschreibung: Beiträge zum Verstehen kultureller Systeme*. *Suhrkamp-Taschenbuch Wissenschaft: Vol. 696*. Frankfurt am Main: Suhrkamp.

GfK. (2009). *Amazon top of the online shopping websites: Press release of the "Gesellschaft für Konsumforschung"*. Retrieved from http://www.gfk.com/group/press_information/press_releases/003904/index.en.html

Gold, L. (2008). Online-Forschung in den USA boomt. *BVM inbrief*, pp. 46–48.

González, V. M., & Mark, G. (2004). Constant, constant, multi-tasking craziness: Managing multiple working spheres. In: *Proceedings of the CHI '04 conference* (pp. 113-120). New York, NY: ACM.

González-Teruel, A., Abad-García, M. F., Sanjuan-Nebot, L., Campón-Gozalvo, J. & Castillo-Blasco, L. (2004). Uso de internet por los médicos colegiados de Valencia: un estudio de viabilidad de la Biblioteca Médica Virtual del Colegio Oficial de Médicos de Valencia. *El Profesional de la Información*, *13*(2), 100–106.

Greifeneder, E. (2010). A content analysis on the use of methods in online user research. In: I. Verheul, A. M. Tammaro & S. Witt (Eds.), *Digital Library Future. User perspectives and institutional strategies. IFLA Publications: Vol. 146.* Munich: K.G.Saur.

Greifeneder, E. (2011a). Einführung in die Online-Benutzerforschung zu Digitalen Bibliotheken. In: B. Bekavac, R. Schneider & W. Schweibenz (Eds.), *Benutzerorientierte Bibliotheken im Web* (pp.75–94). Berlin: De Gruyter Saur.

Greifeneder, E. (2011b). The impact of distraction in natural environments on user experience research. In: *Proceedings of the TPDL'11 Lecture Notes on Computer Science* (pp.308-315). Berlin, Heidelberg: Springer-Verlag.

Griffiths, J.-M. & King, D. (2007). Physical spaces and virtual visitors: The methodologies of comprehensive study of users and uses of museums. In: J. Trant & D. Bearman (Eds.), *International cultural heritage informatics meeting (ICHIM07).* Toronto: Archives & Museum Informatics.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1).

Haefele, C. & Ray, L. (2011). *Using virtual focus groups in distance learning & online environments*. LITA. Retrieved from http://connect.ala.org/node/137552

Hargittai, E. (2002). Beyond logs and surveys: in-depth measures of people's web use skills. *Journal of the American Society for Information Science and technology*, *53*(14), 1239–1244.

Head, A. J. & Eisenberg, M. B. (2011). *Project information literacy research report: "balancing act"*. Retrieved from http://projectinfolit.org/pdfs/PIL_Fall2011_TechStudy_FullReport1.1.pdf

Hilligoss, B. & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: construct, heuristics, and interaction in context. *Information Processing & Management*, *44*(4), 1467–1484.

Hine, C. (2006). *Virtual methods: Issues in social research on the internet*. Oxford: Berg.

Hine, C., Kendall, L. & boyd, d. (2009). How can qualitative internet researchers define the boundaries of their projects? In: A. N. Markham & N. K. Baym (Eds.), *Internet inquiry. Conversations about method.* Los Angeles, CA: Sage Publications.

Hodge, V. & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*, 85–126.

Huang, S.-C., Bias, R. G., Payne, T. L. & Rogers, J. B. (2009). Remote usability testing: a practice. In: *Proceedings of the JCDL 2009 conference.* (p.397). New York, NY: ACM.

Humboldt-Universität zu Berlin. (2012). *Facts and figures*. Retrieved from http://www.hu-berlin.de/ueberblick-en/humboldt-universitaet-zu-berlin-en/facts/

Hussain, Z. & Griffiths, M. D. (2009). The attitudes, feelings, and experiences of online gamers: a qualitative analysis. *Cyberpsychology & Behavior*, *12*(6), 747–753.

IRN Research. (2011). *Europeana – online visitor survey: Europeana online survey report 2011*. Birmingham, UK. Retrieved from http://version1.europeana.eu/

ISO 9241-210: Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems. (2010). Genf: ISO.

Johansen, S. A., San Agustin, J., Skovsgaard, H., Hansen, J. P. & Tall, M. (2011). Low cost vs. high-end eye tracking for usability testing. In: *Proceedings of the CHI '11 conference* (pp.1177-1182). New York, NY: ACM.

Julien, H., Pecoskie, J. & Reed, K. (2011). Trends in information behavior research, 1999–2008: A content analysis. *Library & Information Science Research*, *33*, 19–24.

Kazmer, M. M. & Xie, B. (2008). Qualitative interviewing in internet studies: Playing with the media, playing with the method. *Information, Communication & Society*, *11*(2), 257–278.

Kellar, M., Watters, C., Duffy, J., & Shepherd, M. (2004). Effect of task on time spent reading as an implicit measure of interest. *Proceedings of the American Society for Information Science and Technology*, *41*(1), 168–175.

Kelly, D. (2006a). Measuring online information seeking context, Part 1: Background and method. *Journal of the American Society for Information Science and Technology*, *57(14)*, 1729-1739.

Kelly, D. (2006b). Measuring online information seeking context, Part 2: Findings and discussion. *Journal of the American Society for Information Science and Technology*, *57(14)*, 1862-1874.

Kelly, D. & Gyllstrom, K. (2011). An examination of two delivery modes for interactive search system experiments: remote and laboratory. In: *Proceedings of the CHI ’11 conference* (pp.1531–1540). New York, NY: ACM.

Kelly, D., Harper, D. J. & Landau, B. (2008). Questionnaire mode effects in interactive information retrieval experiments. *Information Processing and Management*, *44*(1), 122-141.

Kirschner, P. A. & Karpinski, A. C. (2010). Facebook® and academic performance. *Computers in Human Behavior*, *26*(6), 1237-1245.

Kittur, A., Chi, E. H. & Bongwon, S. (2008). Crowdsourcing user studies with Mechanical Turk. In: *Proceedings of the CHI ’08 conference* (pp.453–456). New York, NY: ACM.

Krug, S. & Dubau, J. (2006). *Don't make me think!: Web Usability – das intuitive Web* (2. Aufl.). Bonn: mitp.

Law, A. S., Logie, R. H. & Pearson, D. G. (2006). The impact of secondary tasks on multitasking in a virtual environment. *Acta Psychologica*, *122*(1), 27–44.

Lee, J.-J. & Lee, K.-P. (2007). Cultural differences and design methods for user experience research: Dutch and Korean participants compared. In: I. Koskinen & K. Turkka (Eds.), *International conference on designing pleasurable products and interfaces* (pp.20–34). Helsinki, Finland.

Lim, S. & Simon, C. (2011). Credibility judgment and verification behavior of college students concerning Wikipedia. *First Monday*, *16*(4).

Liu, Z. (2006). Print vs. electronic resources: a study of user perceptions, preferences, and use. *Information Processing & Management*, *42*(2), 583–592.

Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., Zhang, J., Zhang, X. (2010). Search behaviors in different task types. In: *Proceedings of the JCDL’10 conference* (pp. 69–78). New York, NY: ACM.

Lyons, R. (2011). Statistical correctness. *Library & Information Science Research*, *33*, 92–95.

Mankoff, J., Fait, H. & Tran, T. (2005). Is your web page accessible?: A comparative study of methods for assessing web page accessibility for the blind. In: *Proceedings of the CHI ’05 conference* (pp. 41–50). New York, NY: ACM.

Mark, G., Gudith, D., & Klocke, U. (2008). The cost of interrupted work: More speed and stress. In: *Proceedings of the CHI ’08 conference* (pp. 107–110). New York, NY: ACM.

Markham, A. N. & Baym, N. K. (2009). Introduction: Making smart choices on shifting ground. In: A. N. Markham & N. K. Baym (Eds.), *Internet inquiry. Conversations about methods* (pp.vii–xvii). Los Angeles, CA: Sage Publications.

McFadden, E., Hager, D. R., Elie, C. J. & Blackwell, J. M. (2002). Remote usability evaluation: overview and case studies. *International Journal on Human Computer Interaction*, *14*(3–4), 489–502.

Meho, L. I. (2006). E-Mail interviewing in qualitative research: A methodological discussion. *Journal of the American Society for Information Science and technology*, *57*(10), 1284–1295.

Miller, D. & Slater, D. (2000). *The Internet: An ethnographic approach*. Oxford: Berg.

Nicholas, D., Huntington, P., Jamali, H. R. & Tenopir, C. (2006). Finding information in (very large) digital libraries: a deep log approach to determining differences in use according to method of access. *Journal of Academic Librarianship*, *32*(2), 119–126.

Nielsen, J. (2000). *Designing Web usability*. Indianapolis, IN: New Riders.

Nielsen, J. (2006). *Outliers and Luck in User Performance*. Jakob Nielsen's Alertbox. Retrieved from http://www.useit.com/alertbox/outlier_performance.html

Nieminen, M. P., Mannonen, P. & Viitanen, J. (2007). International remote usability evaluation: the bliss of not being there. In: N. Aykin (Ed.). Usability and Internationalization. HCII '07. Lecture Notes in Computer Science: Vol. 4559 (pp.388–397). Berlin: Springer.

Orgad, S., Bakardjieva, M. & Gajjala, R. (2009). How can researchers make sense of the issues involved in collecting and interpreting online and offline data? In: A. N. Markham & N. K. Baym (Eds.), *Internet inquiry. Conversations about method.* Los Angeles, CA: Sage Publications.

Perloff, R. M. (2003). The dynamics of persuasion: Communication and attitudes in the 21st century. In: R. M. Perloff (Ed.), *Processing Persuasive Communications* (2nd ed., pp.119–148). Mahwah, NJ: Lawrence Erlbaum.

Petrie, H., Hamilton, F., King, N. & Pavan, P. (2006). Remote usability evaluations with disabled people. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp.1133-1141). New York, NY: ACM.

Petty, R. E. & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In: L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp.123–205). Academic Press.

Polkehn, K., Wandke, H. & Dahm, M. (2010). Usability-Evaluation interaktiver Geräte: Online vs. Labor? In: *Mensch & Computer 2010: Interaktive Kulturen*.

Pomerantz, J., Choemprayons, S. & Eakin, L. (2008). The development and impact of digital library funding in the United States. *Advances in Librarianship*, *31*, 37–92.

Ramm, M., Multrus F. & Bargel, T. (2011). *Studiensituation und studentische Orientierungen: 11. Studierendensurvey an Universitäten und Fachhochschulen, Langfassung*. Bonn, Berlin.

Rieh, S. Y. (2004). On the Web at home: Information seeking and Web searching in the home environment. *Journal of the American Society for Information Science and technology*, *55*(8), 743–753.

Roethlisberger, F. J., Dickson, W. J. & Wright, H. A. (1975). *Management and the worker: An account of a research program conducted by the Western electric Company, Hawthorne Works, Chicago* (16. printing.). Cambridge, MA: Harvard Univ. Press.

Rotman, D., Preece, J., He, Y. & Druin, A. (2012). Extreme ethnography: challenges for research in large scale online environments. In: *Proceedings of the iconference '12* (pp.207-214). New York, NY: ACM.

Rousseau, M., Simon, M., Bertrand, R. & Hachey, K. (2011). Reporting missing data: a study of selected articles published from 2003–2007. *Quality & Quantity*, 1–14.

Ruder, M. & Bless, H. (2003). Mood and the reliance on the ease of retrieval heuristic. *Journal of Personality and Social Psychology*, *85*(1), 20–32.

Rui, Y. & Huang, T. S. (1999). Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, *10*, 39-62.

Rutgers University School of Communication and Information. (2009). *PooDLE: Personalization of Digital Library Experience*. Retrieved from http://comminfo.rutgers.edu/imls/poodle/index.html

Safley, E. (2006). Demand for e-books in an academic library. *Journal of Library Administration*, *45*(3/4), 445–457.

Saracevic, T. (2010). The notion of context in "Information Interaction in Context". In: Proceedings of the *IIiX '10 conference* (pp.1-2). New York, NY: ACM.

Seadle, M. (2000). Project ethnography: an anthropological approach to assessing digital library services. *Library Trends*, *Fall*.

Selvaraj, P. (2004). *Comparative study of synchronous remote and traditional in-lab usability evaluation methods. Master thesis.* Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Sexton, A., Turner, C., Yeo, G. & Hockey, S. (2004). Understanding users: a prerequisite for developing new technologies. *Journal of the Society of Archivists*, *25*(1), 33–49.

Shneiderman, B. & Plaisant, C. (2005). *Designing the user interface: Strategies for effective human-computer interaction*. Boston, MA: Addison-Wesley.

Snow, K., Ballaux, B., Christensen-Dalsgaard, B., Hofman, H., Hofman Hansen, J. & Innocenti, P,. et al. (2008). Considering the user perspective: research into usage and communication of digital information. *D-Lib Magazine*, *May/June*.

Stieger, S. & Göritz, A. (2006). Using Instant Messaging for internet-based interviews. *Cyberpsychology & Behavior*, *9*(5), 552–559.

Symonds, E. (2011). A practical application of SurveyMonkey as a remote usability-testing tool. *Library Hi Tech*, *29*(3), 436–445.

Thompson, K. E., Rozanski, E. P. & Haake, A. R. (2004). Here, there, anywhere: remote usability testing that works. In: *Proceedings of the SIGITE '04. IT education – the state of the art.* New York, NY: ACM.

Troll Covey, D. (2002). *Usage and usability assessment: library practices and concerns*. Washington, DC. Retrieved from http://www.diglib.org/pubs/dlf096/dlf096.htm

Tullis, T., Fleischman, S., McNult, M., Cianchette, C. & Bergel, M. An empirical comparison of lab and remote usability testing of web sites. In: *Proceedings of the Usability Professionals Association conference 2002*.

Vakkari, P. (1999). Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information Processing & Management*, *35*(6), 819–837.

Welker, M. & Wenzel, O. (2010). *Online-Forschung 2007: Grundlagen und Fallstudien*. *Neue Schriften zur Online-Forschung: Vol. 1*. Köln: von Halem.

Wenzel, O. & Hofmann, O. (2005). 10 Jahre Online-Forschung: Von Visionen, dem Boden der Tatsachen und letztlich doch erfüllten Erwartungen. *planung & analyse*, *1*, 24–28.

West, R. & Lehman, K. (2006). Automated summative usability studies: an empirical evaluation. In: *Proceedings of the CHI '06 conference* (pp.631-639). New York, NY: ACM.

Xia, W. (2003). Digital library services: perceptions and expectations of user communities and librarians in a New Zealand academic library. *Australian Academic and Research Libraries*, *34*(1), 56–70.

11.2 Online resources

Adobe Connect: web conferencing software. Retrieved from http://www.adobe.com/products/adobeconnect.html

Amazon. Retrieved from http://www.amazon.de

AttrakDiff: a service of User Interface Design GmbH. Retrieved from http://www.attrakdiff.de

DigiZeitschriften: Das Deutsche Digitale Zeitschriftenarchiv. Retrieved from http://www.digizeitschriften.de/

EasyUsability.com: cheap and highly targeted usability testing. Retrieved from http://easyusability.com/content

Europeana. Retrieved from http://www.europeana.eu

Facebook. Retrieved from http://www.facebook.com/

FlockDraw: free online drawing tool – collaborative group whiteboard. Retrieved from http://flockdraw.com/

LimeSurvey. (2011). Retrieved from http://www.limesurvey.org/

Mikogo – Videoconferencing. Retrieved from http://www.mikogo.de

ORKA: Open Repository Kassel. Retrieved from http://orka.bibliothek.uni-kassel.de/

Perseus Digital Library. Retrieved from http://www.perseus.tufts.edu/hopper/

Project IsoMetrics. (2009). Retrieved from http://www.isometrics.uni-osnabrueck.de/

Questionnaire for user interaction satisfaction. (2002). Retrieved from http://lap.umd.edu/quis/

Loop11: remote & online usability testing tool. Retrieved from http://www.loop11.com/

Moodle der Humboldt-Universität zu Berlin. Retrieved from http://moodle.hu-berlin.de/

SSOAR: Social Science Open Access Repository. Retrieved from http://www.ssoar.info/

SUMI Questionnaire Homepage. Retrieved from http://sumi.ucc.ie/

SurveyMonkey: free online survey software & questionnaire tool. Retrieved from http://www.surveymonkey.com/

The Valley of the Shadow: two communities in the American Civil War. Retrieved from http://valley.lib.virginia.edu/

Twiddla: Painless Team Collaboration for the Web. Retrieved from http://www.twiddla.com/

Usabilla.com: improve your user experience with continuous design feedback. Retrieved from http://usabilla.com/

UserTesting.com: low cost usability testing. (2012). Retrieved from http://www.usertesting.com/

UserZoom: online usability testing. Retrieved from http://www.userzoom.com/

Webnographer. Retrieved from http://www.webnographer.com/

Wimba Classroom. Retrieved from http://www.wimba.com/products/wimba_classroom

All websites were last accessed 4th April 2012.

Appendix 1. List of variables collected during the experiment

Variable groups are categorized in measurements of time, control variable, measurement of task completion and judgments as suggested in the conceptual framework. Additional variables, which were necessary to examine the settings, are listed at the end of the table. Because of formal university regulations, the *SPSS* files are not included in this work. Both can be provided by the author, though. Variable names are only provided if they are used in form of an abbreviated variable name within the work. Variable codes refer to the code in the *SPSS* files and in the correlations in appendix 7.

Variable group	Variable name	Label	Variable code	Measurement
time scores	test duration	time spent on the whole test in seconds	tts	scale
time scores	time on tasks	computed variable, time spent on all tasks	ttstask	scale
time scores	time on questions	computed variable, time spent on the questionnaire	ttsque	scale
time scores	–	average time spent on the whole test	att	scale
time scores	time on the task in Perseus	time spent on the task in *Perseus*	perts	scale
time scores	time on the task in SSOAR	time spent on the task in *SSOAR*	ssoarts	scale
time scores	time on the task in Bundesarchiv	time spent on the task in the *Bundesarchiv*	bundts	scale
time scores	time on the task in Valley of the Shadow	time spent on the task in *Valley of the Shadow*	valts	scale
time scores	time on the task in Amazon	time spent on the task in *Amazon*	amats	scale
control variable	open programs	existence of other open programs during the test (yes /no)	progrgesch1	ordinal
control variable	frequency of programs	frequency of looking at other programs	progrgesch	scale
control variable	–	existence of open programs (yes/no): chat, Facebook, email, music, web browser, diverse software	progrchat progrfacebook progrmail progrmusic progrweb progrdiverse	ordinal
control variable	contacts	existence of contacts during the test (yes/no)	kt	ordinal
control variable	frequency of contacts	frequency of contacts during the test	ktanz	scale
control variable	technical problems	occurrence of a technical problem during the test	problemvor	ordinal

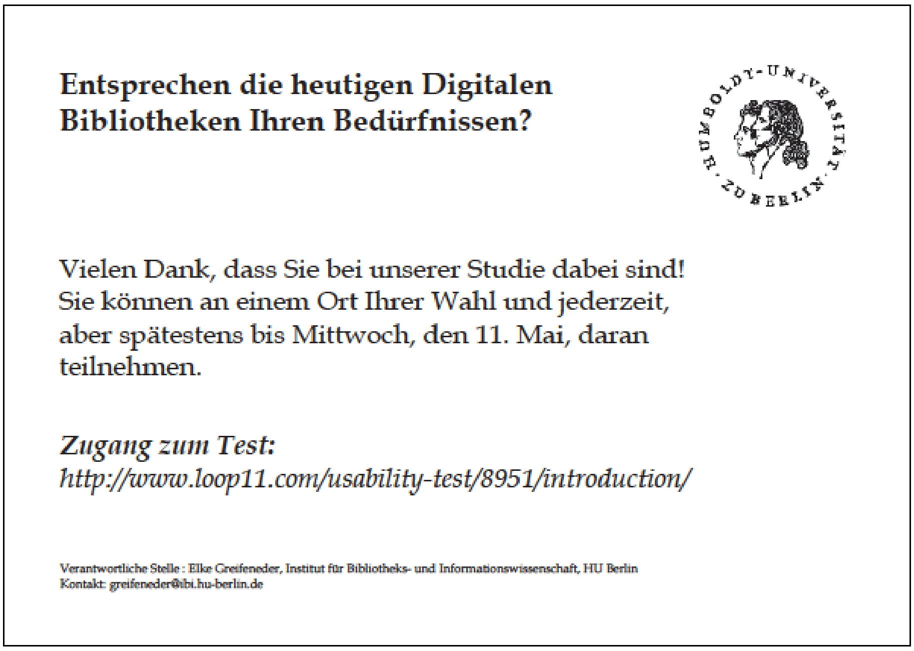
Variable group	Variable name	Label	Variable code	Measurement
control variable	–	description of technical problem	problemtech	nominal
control variable	–	measurement of level of attention: “How much have you been distracted by other programs?” (8-point Likert scale)	prograbgelenkt	ordinal
control variable	–	measurement of level of attention: “In how far have you been distracted by contacts?” (8-point Likert scale)	ktabgelenkt	ordinal
control variable	daydreams	measurement of internal distraction: How much have you been distracted by distracting thoughts like daydreams? (8-point Likert scale)	tagtraum	ordinal
control variable	–	hour of test completion	start	ordinal
control variable	internet connection	kind of internet connection (choice of five groups)	zugang	ordinal
control variable	age	participants’ age	alter	scale
control variable	gender	participants’ gender	geschlecht	ordinal
control variable	–	existing task related knowledge, operationalized as number of semester at the university (choice of five groups)	semester	ordinal
control variable	number of semesters	existing task related knowledge, operationalized as subgroup of variable semester (three groups: beginners, advanced, experts)	semester2	ordinal
control variable	academic specialization	first academic specialization (organized in three groups)	studiengang	ordinal
control variable	–	second academic subject (organized in three groups)	studiengang2	ordinal
control variable	–	third academic subject (organized in three groups)	studiengang3	ordinal
control variable	university degree	level of education background, operationalized as highest degree (choice of seven)	bildung	ordinal
control variable	–	prior experiences with digital libraries, operationalized and coded by the number of digital libraries a participant already knew in addition to *Amazon* (0–4)	kenntnissedb2	scale
control variable	language skills German	language skills in German (choice of two)	kenntnisse-deutsch	ordinal
control variable	language skills English	language skills in English (choice of six)	kenntnisse-engl	ordinal
control variable	–	measurement of mood: “I feel good” (8-point Likert scale)	stimmgut	ordinal
control variable	–	measurement of mood: “I am unsatisfied” (8-point Likert scale)	stimmnicht-gut	ordinal
control variable	mood	aggregated variable mood (stimmgut and inverted variable stimmnichtgut)	mood	ordinal
task completion	page views	total number of page views on all tasks	tpvtask	scale

Variable group	Variable name	Label	Variable code	Measurement
task completion	–	average number of page views on tasks	apv	scale
task completion	number of successful task completions in Perseus	result of task completion in *Perseus*	persuccess	ordinal
task completion	number of successful task completions in SSOAR	result of task completion in *SSOAR*	ssoarsuccess	ordinal
task completion	number of successful task completions in Bundesarchiv	result of task completion in *Bundesarchiv*	bundsuccess	ordinal
task completion	number of successful task completions in Valley of the Shadow	result of task completion in *Valley of the Shadow*	valsuccess	ordinal
task completion	number of successful task completions in Amazon	result of task completion in *Amazon*	amasuccess	ordinal
judgments	–	participants' general evaluation of the digital library *Perseus* (8-point Likert scale)	perges	ordinal
judgments	–	participants' general evaluation of the digital library *SSOAR* (8-point Likert scale)	ssoarges	ordinal
judgments	–	participants' general evaluation of the digital library *Bundesarchiv* (8-point Likert scale)	bundges	ordinal
judgments	–	participants' general evaluation of the digital library *Valley of the Shadow* (8-point Likert scale)	valges	ordinal
judgments	–	participants' general evaluation of the website *Amazon* (8-point Likert scale)	amages	ordinal
judgments	–	perceived difficulty of task in *Perseus* ("this task was difficult"; 8-point Likert scale)	perdif	ordinal
judgments	–	perceived difficulty of task in *SSOAR* (8-point Likert scale)	ssoardif	ordinal
judgments	–	perceived difficulty of task in *Bundesarchiv* (8-point Likert scale)	bunddif	ordinal
judgments	–	perceived difficulty of task in *Valley of the Shadow* (8-point Likert scale)	valdif	ordinal
judgments	–	perceived difficulty of task in *Amazon* (8-point Likert scale)	amadif	ordinal
judgments	–	perceived relevance of search result in *Perseus* (8-point Likert scale)	perarg1	ordinal
judgments	–	perceived relevance of search result in *SSOAR* (8-point Likert scale)	ssoararg1	ordinal

Variable group	Variable name	Label	Variable code	Measurement
judgments	–	perceived relevance of search result in *Bundesarchiv* (8-point Likert scale)	bundarg1	ordinal
judgments	–	perceived relevance of search result in *Valley of the Shadow* (8-point Likert scale)	valarg1	ordinal
judgments	–	perceived relevance of search result in *Amazon* (8-point Likert scale)	amarg1	ordinal
judgments	–	rating of search functionality in *Perseus* (8-point Likert scale)	perarg2	ordinal
judgments	–	rating of search functionality in *SSOAR* (8-point Likert scale)	ssoararg2	ordinal
judgments	–	rating of search functionality in *Bundesarchiv* (8-point Likert scale)	bundarg2	ordinal
judgments	–	rating of search functionality in *Valley of the Shadow* (8-point Likert scale)	valarg2	ordinal
judgments	–	rating of search functionality in *Amazon* (8-point Likert scale)	amarg2	ordinal
judgments	–	rating of the design of the digital library *Perseus* (8-point Likert scale)	percue1	ordinal
judgments	–	rating of the design of the digital library *SSOAR* (8-point Likert scale)	ssoarcue1	ordinal
judgments	–	rating of the design of the digital library *Bundesarchiv* (8-point Likert scale)	bundcue1	ordinal
judgments	–	rating of the design of the digital library *Valley of the Shadow* (8-point Likert scale)	valcue1	ordinal
judgments	–	rating of the design of the website *Amazon* (8-point Likert scale)	amacue1	ordinal
judgments	–	rating of the professional appearance of the digital library *Perseus* (8-point Likert scale)	percue2	ordinal
judgments	–	rating of the professional appearance of the digital library *SSOAR* (8-point Likert scale)	ssoarcue2	ordinal
judgments	–	rating of the professional appearance of the digital library *Bundesarchiv* (8-point Likert scale)	bundcue2	ordinal
judgments	–	rating of the professional appearance of the digital library *Valley of the Shadow* (8-point Likert scale)	valcue2	ordinal
judgments	–	rating of the professional appearance of the digital library *Amazon* (8-point Likert scale)	amacue2	ordinal
judgments	arguments in task Perseus	aggregated variable that measured the “arguments” in *Perseus*	perarg	ordinal
judgments	arguments in task SSOAR	aggregated variable that measured the “arguments” in *SSOAR*	ssoararg	ordinal
judgments	arguments in task Bundesarchiv	aggregated variable that measured the “arguments” in *Bundesarchiv*	bundarg	ordinal

Variable group	Variable name	Label	Variable code	Measurement
judgments	arguments in task Valley of the Shadow	aggregated variable that measured the "arguments" in *Valley of the Shadow*	valarg	ordinal
judgments	arguments in task Amazon	aggregated variable that measured the "arguments" in *Amazon*	amarg	ordinal
judgments	cues in task Perseus	aggregated variable that measured the "cues" in *Perseus*	percue	ordinal
judgments	cues in task SSOAR	aggregated variable that measured the "cues" in *SSOAR*	ssoarcue	ordinal
judgments	cues in task Bundesarchiv	aggregated variable that measured the "cues" in *Bundesarchiv*	bundcue	ordinal
judgments	cues in task Valley of the Shadow	aggregated variable that measured the "cues" in *Valley of the Shadow*	valcue	ordinal
judgments	cues in task Amazon	aggregated variable that measured the "cues" in *Amazon*	amacue	ordinal
judgments	cues	cumulated variable that measured the "cues" across all digital libraries using a Fisher transformation	cues	ordinal
judgments	perceived difficulty	cumulated variable that measured the perceived difficulty of task across all digital libraries using a Fisher transformation	dif	ordinal
judgments	general evaluation	cumulated variable that measured participants' general evaluation of the sites across all digital libraries using a Fisher transformation	genev	ordinal
experimental setting	–	participant number	no	scale
experimental setting	settings	setting (laboratory/natural environment)	ort	ordinal
experimental setting	location	detailed setting in the natural environment (grouped in seven choices)	ort2	ordinal

Appendix 2. Examples of recruiter sheets



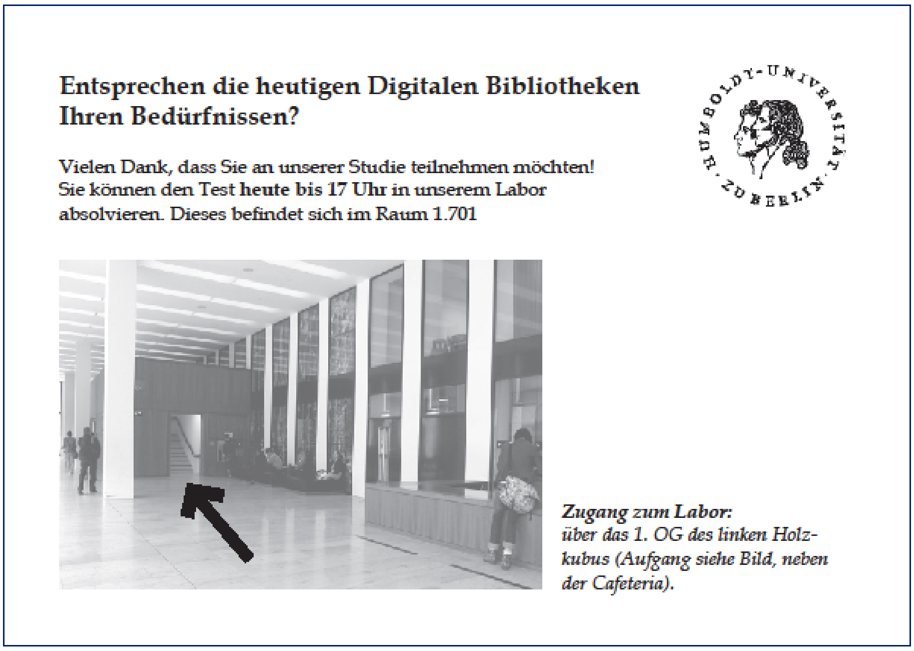
Entsprechen die heutigen Digitalen Bibliotheken Ihren Bedürfnissen?

Vielen Dank, dass Sie bei unserer Studie dabei sind! Sie können an einem Ort Ihrer Wahl und jederzeit, aber spätestens bis Mittwoch, den 11. Mai, daran teilnehmen.

Zugang zum Test:
http://www.loop11.com/usability-test/8951/introduction/

Verantwortliche Stelle : Elke Greifeneder, Institut für Bibliotheks- und Informationswissenschaft, HU Berlin
Kontakt: greifeneder@ibi.hu-berlin.de

Figure 13. Example of a recruiter's sheet for participants in the natural environment.



Entsprechen die heutigen Digitalen Bibliotheken Ihren Bedürfnissen?

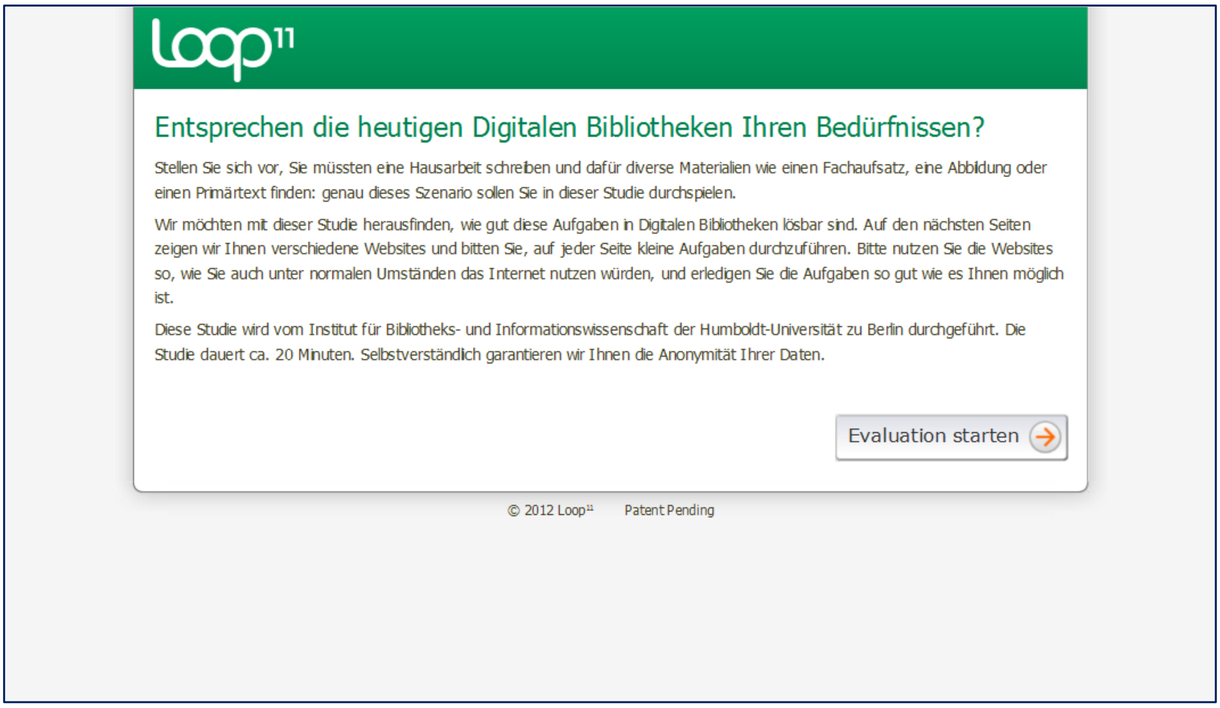
Vielen Dank, dass Sie an unserer Studie teilnehmen möchten! Sie können den Test **heute** bis 17 Uhr in unserem Labor absolvieren. Dieses befindet sich im Raum 1.701

Zugang zum Labor:
über das 1. OG des linken Holzkubus (Aufgang siehe Bild, neben der Cafeteria).

Figure 14. Example of a recruiter's sheet for participants in the laboratory.

Appendix 3. Screenshots of the experiment

The order of the screenshots reflects the original page order in the online test.



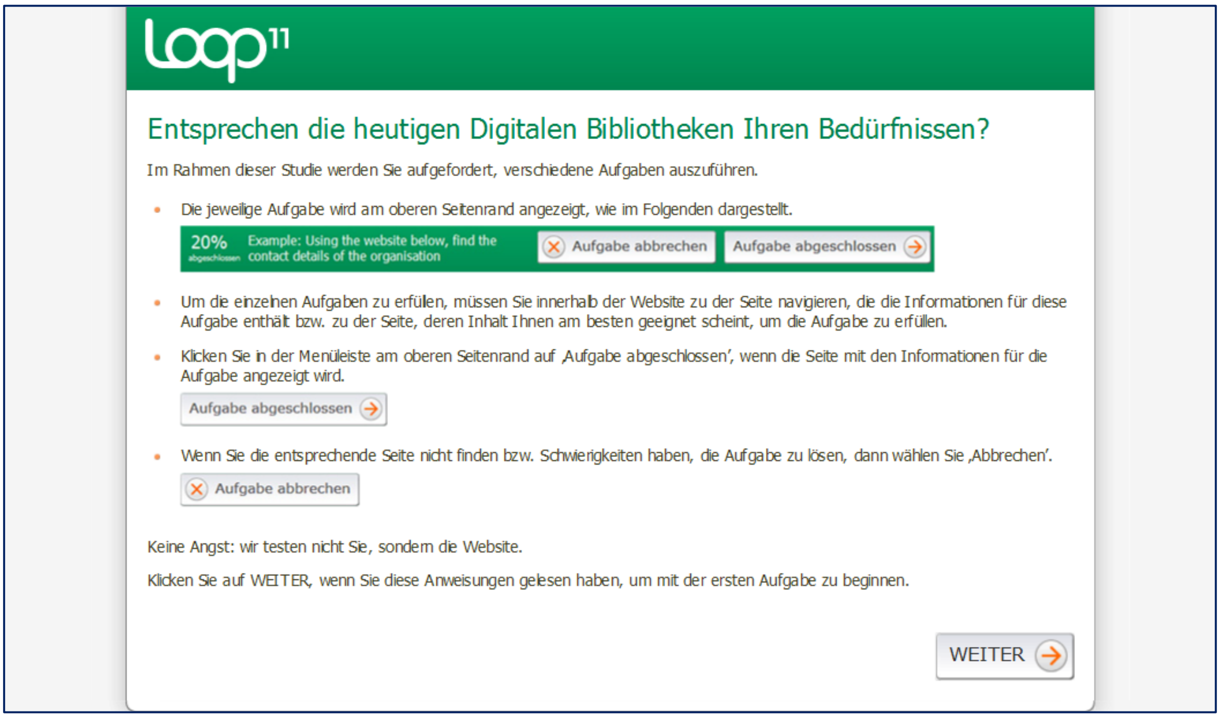
Entsprechen die heutigen Digitalen Bibliotheken Ihren Bedürfnissen?

Stellen Sie sich vor, Sie müssten eine Hausarbeit schreiben und dafür diverse Materialien wie einen Fachaufsatz, eine Abbildung oder einen Primärtext finden: genau dieses Szenario sollen Sie in dieser Studie durchspielen.

Wir möchten mit dieser Studie herausfinden, wie gut diese Aufgaben in Digitalen Bibliotheken lösbar sind. Auf den nächsten Seiten zeigen wir Ihnen verschiedene Websites und bitten Sie, auf jeder Seite kleine Aufgaben durchzuführen. Bitte nutzen Sie die Websites so, wie Sie auch unter normalen Umständen das Internet nutzen würden, und erledigen Sie die Aufgaben so gut wie es Ihnen möglich ist.

Diese Studie wird vom Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin durchgeführt. Die Studie dauert ca. 20 Minuten. Selbstverständlich garantieren wir Ihnen die Anonymität Ihrer Daten.

Figure 15. Welcome page.



Entsprechen die heutigen Digitalen Bibliotheken Ihren Bedürfnissen?

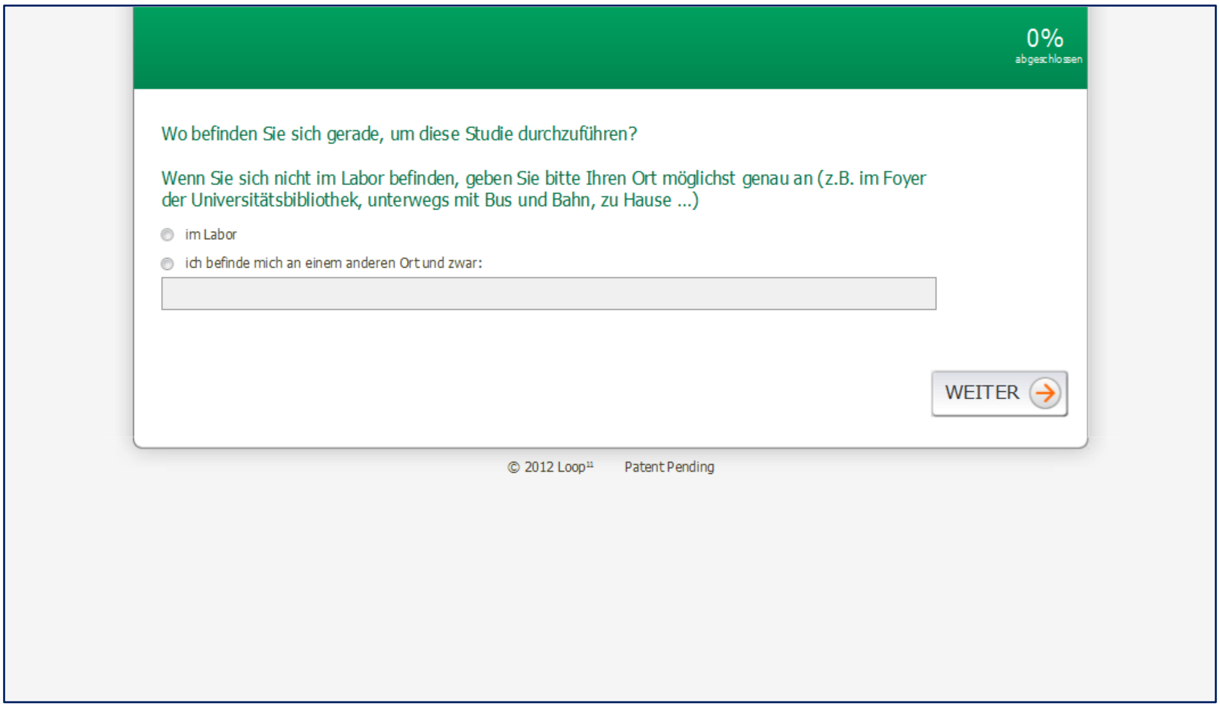
Im Rahmen dieser Studie werden Sie aufgefordert, verschiedene Aufgaben auszuführen.

- Die jeweilige Aufgabe wird am oberen Seitenrand angezeigt, wie im Folgenden dargestellt.
- Um die einzelnen Aufgaben zu erfüllen, müssen Sie innerhalb der Website zu der Seite navigieren, die die Informationen für diese Aufgabe enthält bzw. zu der Seite, deren Inhalt Ihnen am besten geeignet scheint, um die Aufgabe zu erfüllen.
- Klicken Sie in der Menüleiste am oberen Seitenrand auf ‚Aufgabe abgeschlossen', wenn die Seite mit den Informationen für die Aufgabe angezeigt wird.
- Wenn Sie die entsprechende Seite nicht finden bzw. Schwierigkeiten haben, die Aufgabe zu lösen, dann wählen Sie ‚Abbrechen'.

Keine Angst: wir testen nicht Sie, sondern die Website.

Klicken Sie auf WEITER, wenn Sie diese Anweisungen gelesen haben, um mit der ersten Aufgabe zu beginnen.

Figure 16. Help page about test functionality.



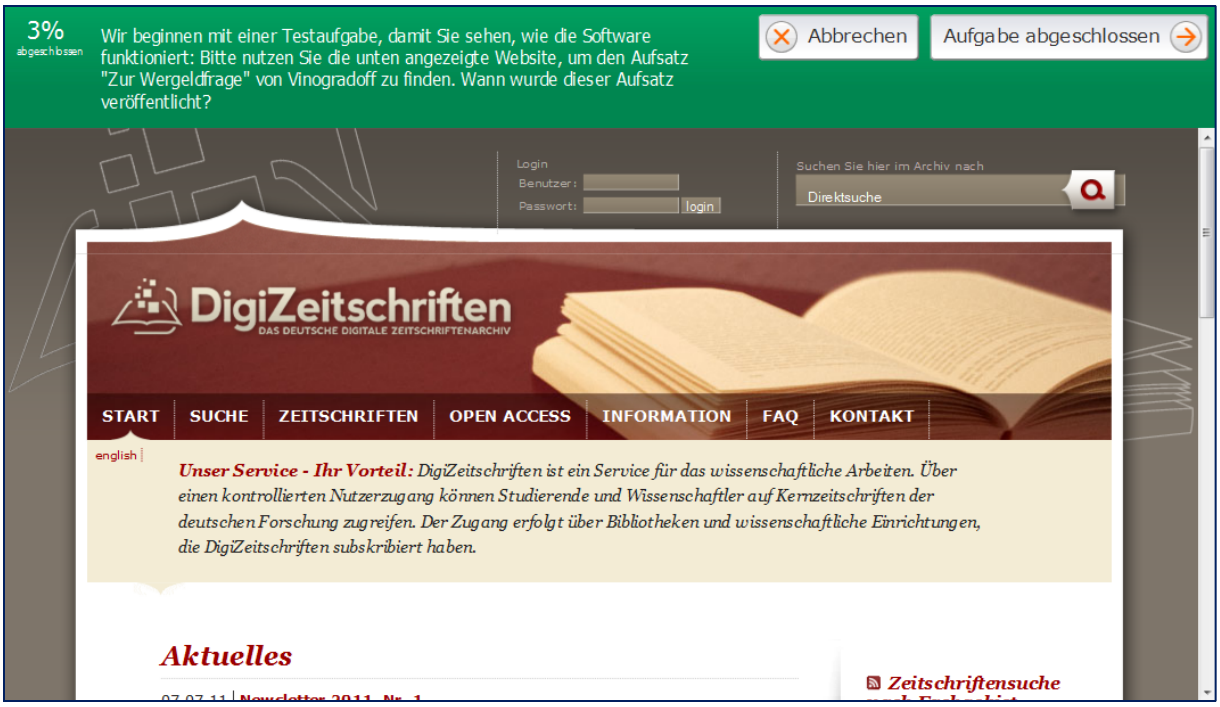
Wo befinden Sie sich gerade, um diese Studie durchzuführen?

Wenn Sie sich nicht im Labor befinden, geben Sie bitte Ihren Ort möglichst genau an (z.B. im Foyer der Universitätsbibliothek, unterwegs mit Bus und Bahn, zu Hause ...)

○ im Labor

○ ich befinde mich an einem anderen Ort und zwar:

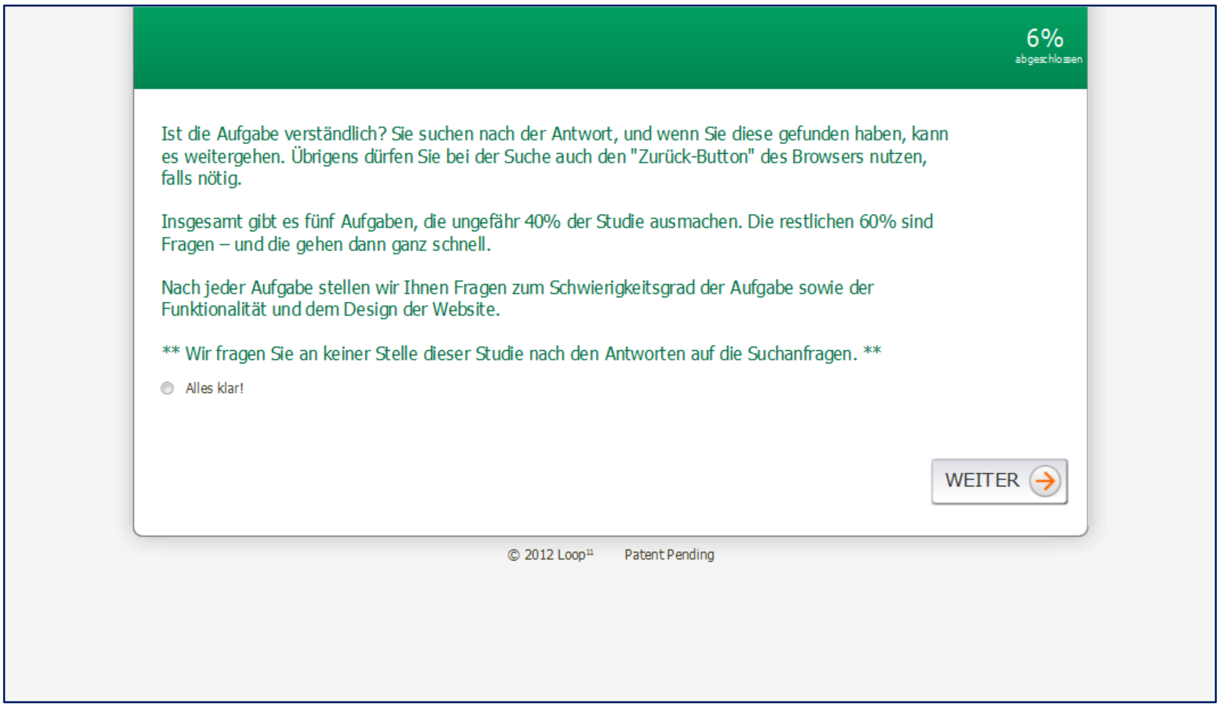
Figure 17. Question about the setting.



Wir beginnen mit einer Testaufgabe, damit Sie sehen, wie die Software funktioniert: Bitte nutzen Sie die unten angezeigte Website, um den Aufsatz "Zur Wergeldfrage" von Vinogradoff zu finden. Wann wurde dieser Aufsatz veröffentlicht?

Unser Service - Ihr Vorteil: DigiZeitschriften ist ein Service für das wissenschaftliche Arbeiten. Über einen kontrollierten Nutzerzugang können Studierende und Wissenschaftler auf Kernzeitschriften der deutschen Forschung zugreifen. Der Zugang erfolgt über Bibliotheken und wissenschaftliche Einrichtungen, die DigiZeitschriften subskribiert haben.

Figure 18. Task in the digital library *DigiZeitschriften* to learn the software and to get accustomed to the kind of tasks.


6%
abgeschlossen

Ist die Aufgabe verständlich? Sie suchen nach der Antwort, und wenn Sie diese gefunden haben, kann es weitergehen. Übrigens dürfen Sie bei der Suche auch den "Zurück-Button" des Browsers nutzen, falls nötig.

Insgesamt gibt es fünf Aufgaben, die ungefähr 40% der Studie ausmachen. Die restlichen 60% sind Fragen – und die gehen dann ganz schnell.

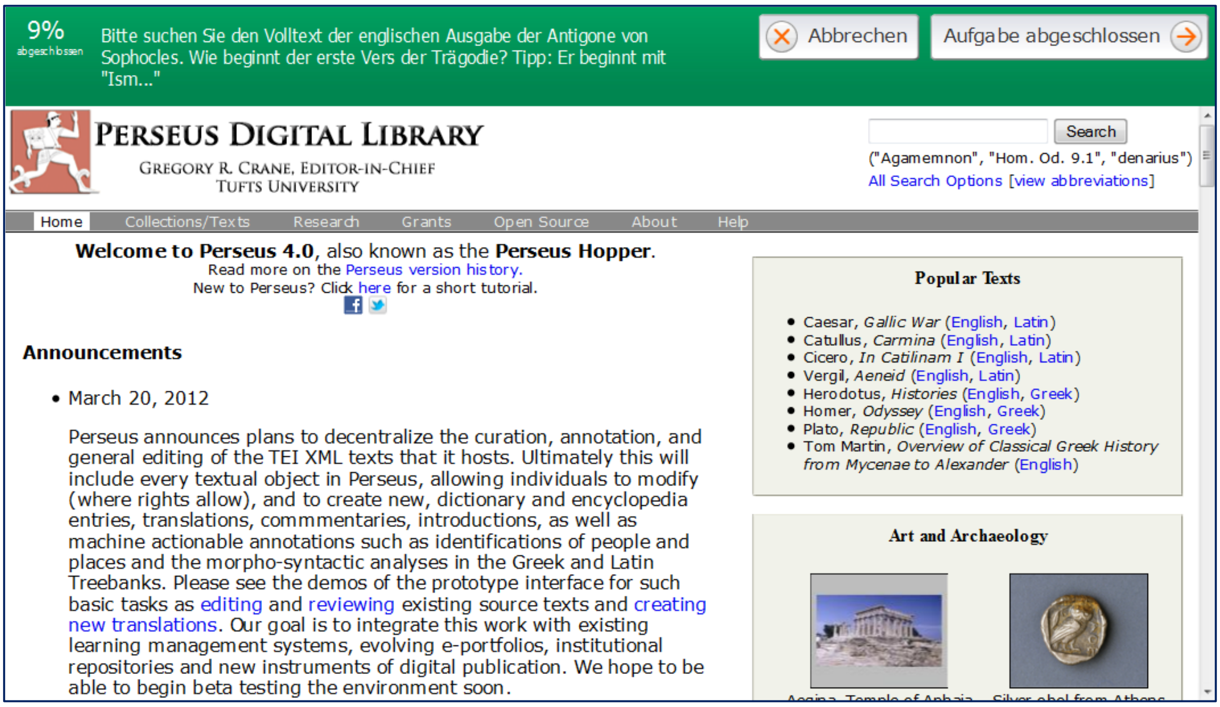
Nach jeder Aufgabe stellen wir Ihnen Fragen zum Schwierigkeitsgrad der Aufgabe sowie der Funktionalität und dem Design der Website.

** Wir fragen Sie an keiner Stelle dieser Studie nach den Antworten auf die Suchanfragen. **

○ Alles klar!

© 2012 Loop11 Patent Pending

Figure 19. Further explanations about the test and the responses to the tasks.


9%
abgeschlossen

Bitte suchen Sie den Volltext der englischen Ausgabe der Antigone von Sophocles. Wie beginnt der erste Vers der Trägodie? Tipp: Er beginnt mit "Ism..."

Welcome to Perseus 4.0, also known as the **Perseus Hopper**.
Read more on the Perseus version history.
New to Perseus? Click here for a short tutorial.

Announcements

- March 20, 2012

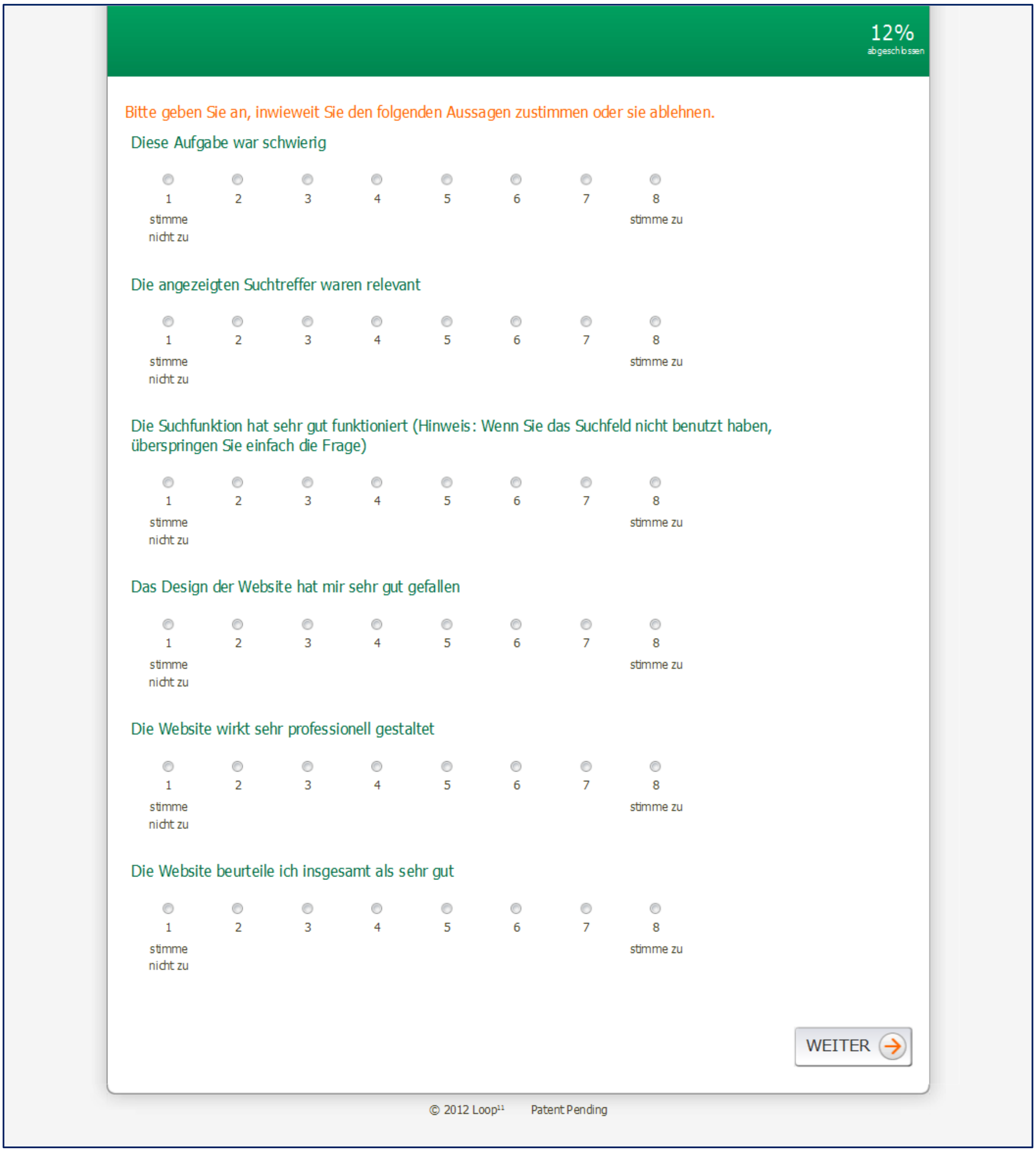
Perseus announces plans to decentralize the curation, annotation, and general editing of the TEI XML texts that it hosts. Ultimately this will include every textual object in Perseus, allowing individuals to modify (where rights allow), and to create new, dictionary and encyclopedia entries, translations, commmentaries, introductions, as well as machine actionable annotations such as identifications of people and places and the morpho-syntactic analyses in the Greek and Latin Treebanks. Please see the demos of the prototype interface for such basic tasks as editing and reviewing existing source texts and creating new translations. Our goal is to integrate this work with existing learning management systems, evolving e-portfolios, institutional repositories and new instruments of digital publication. We hope to be able to begin beta testing the environment soon.

Popular Texts

- Caesar, *Gallic War* (English, Latin)
- Catullus, *Carmina* (English, Latin)
- Cicero, *In Catilinam I* (English, Latin)
- Vergil, *Aeneid* (English, Latin)
- Herodotus, *Histories* (English, Greek)
- Homer, *Odyssey* (English, Greek)
- Plato, *Republic* (English, Greek)
- Tom Martin, *Overview of Classical Greek History from Mycenae to Alexander* (English)

Art and Archaeology

Figure 20. First task in Perseus.



12%
abgeschlossen

Bitte geben Sie an, inwieweit Sie den folgenden Aussagen zustimmen oder sie ablehnen.

Diese Aufgabe war schwierig

○	○	○	○	○	○	○	○
1	2	3	4	5	6	7	8
stimme nicht zu							stimme zu

Die angezeigten Suchtreffer waren relevant

○	○	○	○	○	○	○	○
1	2	3	4	5	6	7	8
stimme nicht zu							stimme zu

Die Suchfunktion hat sehr gut funktioniert (Hinweis: Wenn Sie das Suchfeld nicht benutzt haben, überspringen Sie einfach die Frage)

○	○	○	○	○	○	○	○
1	2	3	4	5	6	7	8
stimme nicht zu							stimme zu

Das Design der Website hat mir sehr gut gefallen

○	○	○	○	○	○	○	○
1	2	3	4	5	6	7	8
stimme nicht zu							stimme zu

Die Website wirkt sehr professionell gestaltet

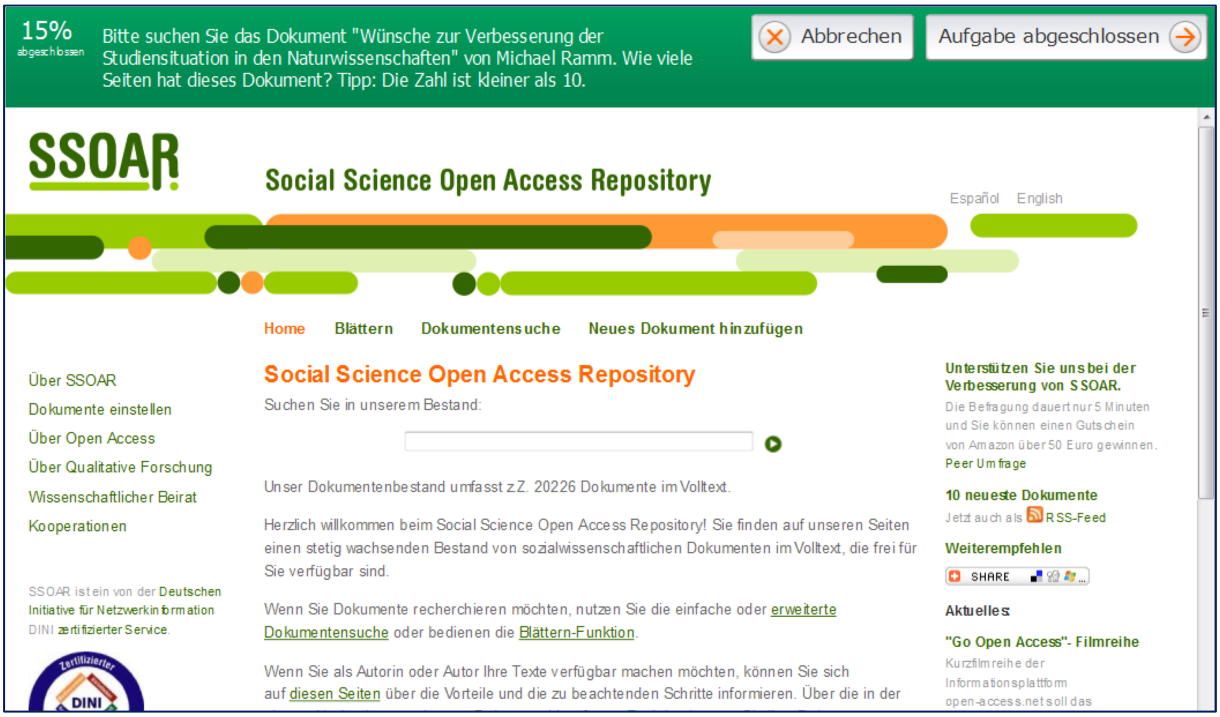
○	○	○	○	○	○	○	○
1	2	3	4	5	6	7	8
stimme nicht zu							stimme zu

Die Website beurteile ich insgesamt als sehr gut

○	○	○	○	○	○	○	○
1	2	3	4	5	6	7	8
stimme nicht zu							stimme zu

© 2012 Loop^{11} Patent Pending

Figure 21. Set of questions after each task.



Bitte suchen Sie das Dokument "Wünsche zur Verbesserung der Studiensituation in den Naturwissenschaften" von Michael Ramm. Wie viele Seiten hat dieses Dokument? Tipp: Die Zahl ist kleiner als 10.

Social Science Open Access Repository

Suchen Sie in unserem Bestand:

Unser Dokumentenbestand umfasst z.Z. 20226 Dokumente im Volltext.

Herzlich willkommen beim Social Science Open Access Repository! Sie finden auf unseren Seiten einen stetig wachsenden Bestand von sozialwissenschaftlichen Dokumenten im Volltext, die frei für Sie verfügbar sind.

Wenn Sie Dokumente recherchieren möchten, nutzen Sie die einfache oder erweiterte Dokumentensuche oder bedienen die Blättern-Funktion.

Wenn Sie als Autorin oder Autor Ihre Texte verfügbar machen möchten, können Sie sich auf diesen Seiten über die Vorteile und die zu beachtenden Schritte informieren. Über die in der

SSOAR ist ein von der Deutschen Initiative für Netzwerkinformation DINI zertifizierter Service.

Unterstützen Sie uns bei der Verbesserung von SSOAR.

Die Befragung dauert nur 5 Minuten und Sie können einen Gutschein von Amazon über 50 Euro gewinnen. Peer Umfrage

10 neueste Dokumente

Jetzt auch als RSS-Feed

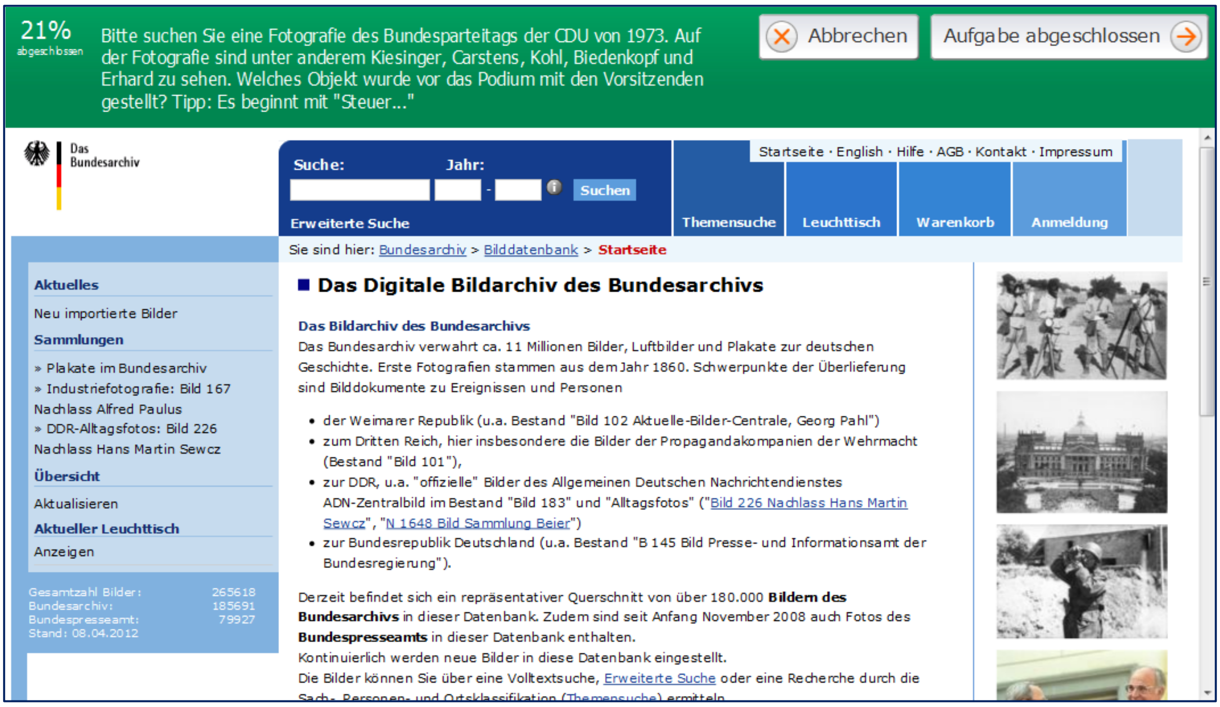
Weiterempfehlen

Aktuelles

"Go Open Access"- Filmreihe

Kurzfilmreihe der Informationsplattform open-access.net soll das

Figure 22. Second task in SSOAR.


21%
abgeschlossen
Abbrechen
Aufgabe abgeschlossen
Das Bundesarchiv
Suche:
Jahr:
Suchen
Erweiterte Suche
Themensuche
Leuchttisch
Warenkorb
Anmeldung
Startseite · English · Hilfe · AGB · Kontakt · Impressum
Sie sind hier: Bundesarchiv > Bilddatenbank > Startseite
Aktuelles
Neu importierte Bilder
Sammlungen
» Plakate im Bundesarchiv
» Industriefotografie: Bild 167 Nachlass Alfred Paulus
» DDR-Alltagsfotos: Bild 226 Nachlass Hans Martin Sewcz
Übersicht
Aktualisieren
Aktueller Leuchttisch
Anzeigen

Bitte suchen Sie eine Fotografie des Bundesparteitags der CDU von 1973. Auf der Fotografie sind unter anderem Kiesinger, Carstens, Kohl, Biedenkopf und Erhard zu sehen. Welches Objekt wurde vor das Podium mit den Vorsitzenden gestellt? Tipp: Es beginnt mit "Steuer..."

Gesamtzahl Bilder:	265618
Bundesarchiv:	185691
Bundespresseamt:	79927
Stand: 08.04.2012	

Das Digitale Bildarchiv des Bundesarchivs

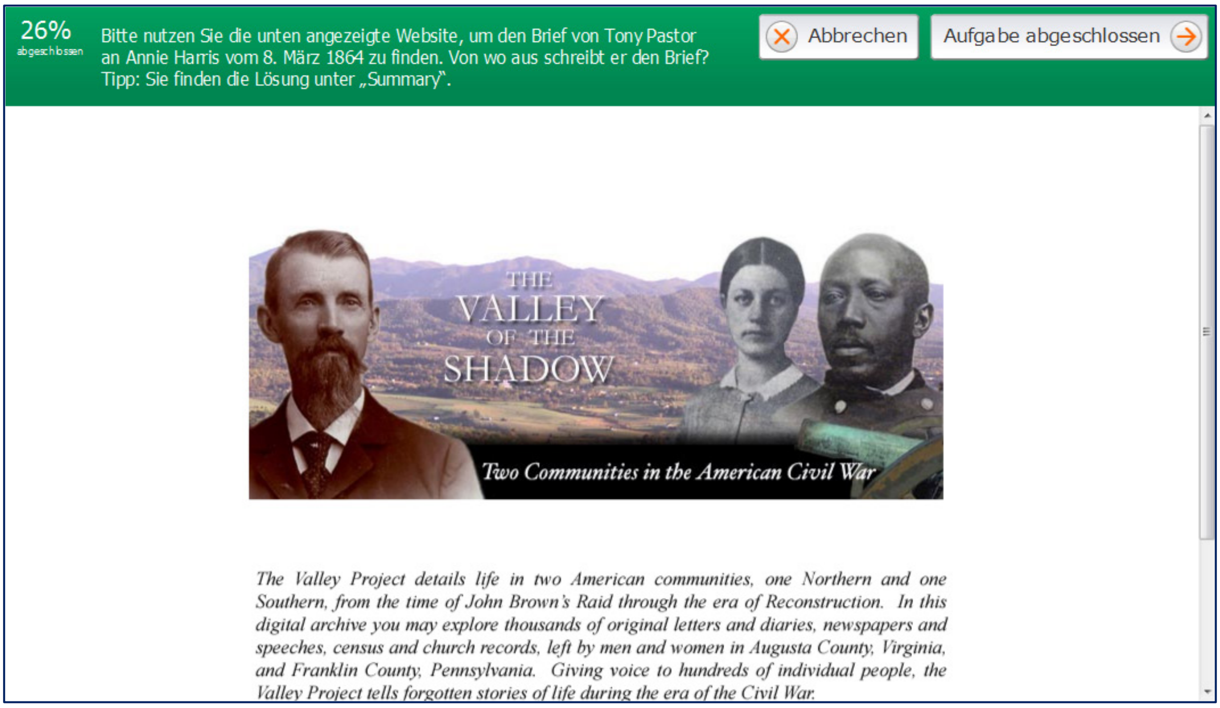
Das Bildarchiv des Bundesarchivs

Das Bundesarchiv verwahrt ca. 11 Millionen Bilder, Luftbilder und Plakate zur deutschen Geschichte. Erste Fotografien stammen aus dem Jahr 1860. Schwerpunkte der Überlieferung sind Bilddokumente zu Ereignissen und Personen

- der Weimarer Republik (u.a. Bestand "Bild 102 Aktuelle-Bilder-Centrale, Georg Pahl")
- zum Dritten Reich, hier insbesondere die Bilder der Propagandakompanien der Wehrmacht (Bestand "Bild 101"),
- zur DDR, u.a. "offizielle" Bilder des Allgemeinen Deutschen Nachrichtendienstes ADN-Zentralbild im Bestand "Bild 183" und "Alltagsfotos" ("Bild 226 Nachlass Hans Martin Sewcz", "N 1648 Bild Sammlung Beier")
- zur Bundesrepublik Deutschland (u.a. Bestand "B 145 Bild Presse- und Informationsamt der Bundesregierung").

Derzeit befindet sich ein repräsentativer Querschnitt von über 180.000 **Bildern des Bundesarchivs** in dieser Datenbank. Zudem sind seit Anfang November 2008 auch Fotos des **Bundespresseamts** in dieser Datenbank enthalten.
Kontinuierlich werden neue Bilder in diese Datenbank eingestellt.
Die Bilder können Sie über eine Volltextsuche, Erweiterte Suche oder eine Recherche durch die Sach-, Personen- und Ortsklassifikation (Themensuche) ermitteln.

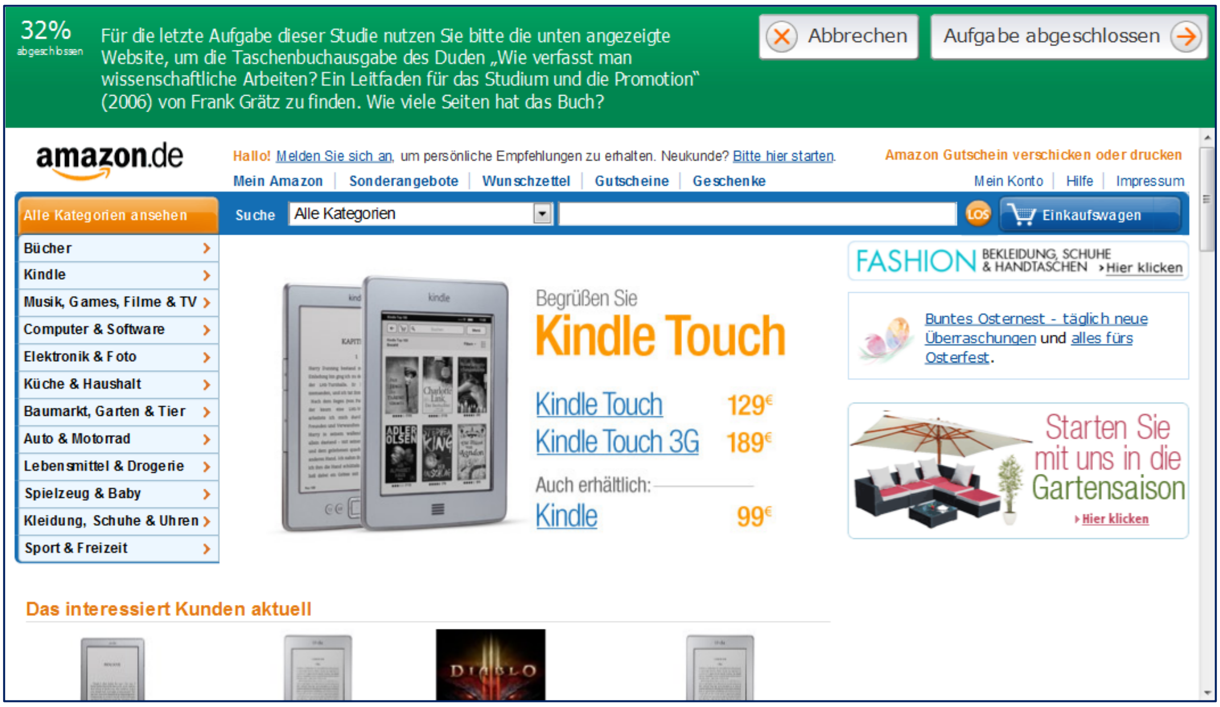
Figure 23. Third task in the Bundesarchiv.



Bitte nutzen Sie die unten angezeigte Website, um den Brief von Tony Pastor an Annie Harris vom 8. März 1864 zu finden. Von wo aus schreibt er den Brief? Tipp: Sie finden die Lösung unter „Summary".

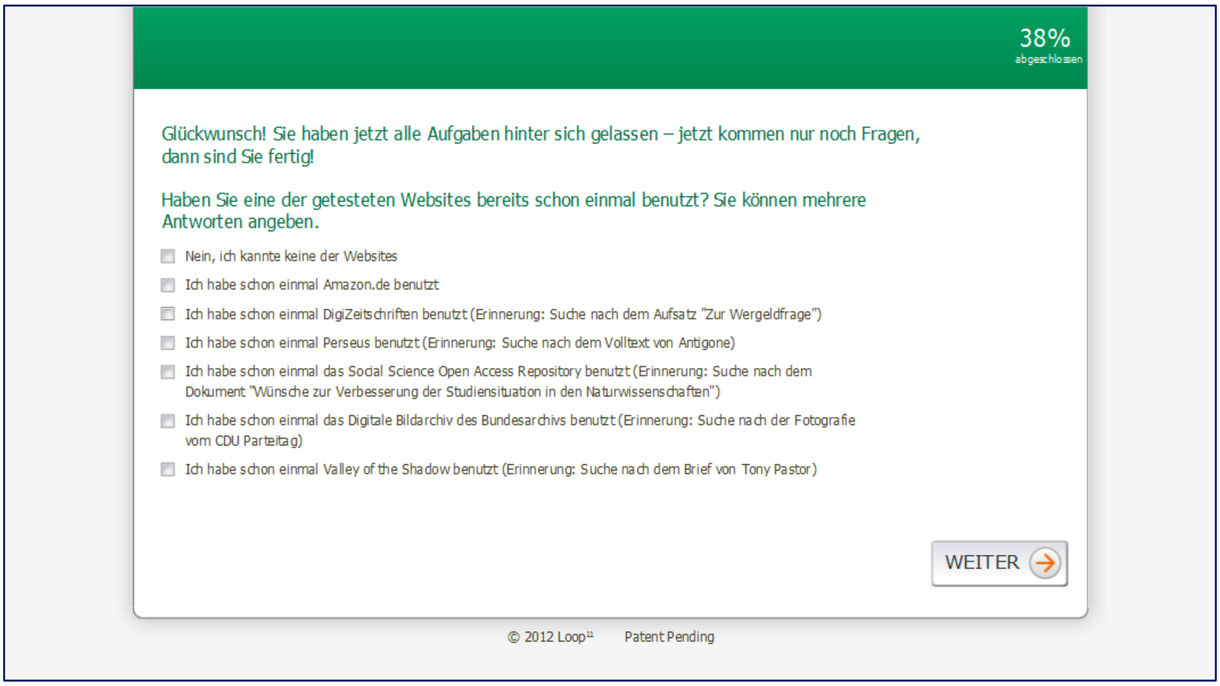
The Valley Project details life in two American communities, one Northern and one Southern, from the time of John Brown's Raid through the era of Reconstruction. In this digital archive you may explore thousands of original letters and diaries, newspapers and speeches, census and church records, left by men and women in Augusta County, Virginia, and Franklin County, Pennsylvania. Giving voice to hundreds of individual people, the Valley Project tells forgotten stories of life during the era of the Civil War.

Figure 24. Fourth task in Valley of the Shadow.



Für die letzte Aufgabe dieser Studie nutzen Sie bitte die unten angezeigte Website, um die Taschenbuchausgabe des Duden „Wie verfasst man wissenschaftliche Arbeiten? Ein Leitfaden für das Studium und die Promotion" (2006) von Frank Grätz zu finden. Wie viele Seiten hat das Buch?

Figure 25. Fifth task in Amazon.



Glückwunsch! Sie haben jetzt alle Aufgaben hinter sich gelassen – jetzt kommen nur noch Fragen, dann sind Sie fertig!

Haben Sie eine der getesteten Websites bereits schon einmal benutzt? Sie können mehrere Antworten angeben.

□ Nein, ich kannte keine der Websites

□ Ich habe schon einmal Amazon.de benutzt

□ Ich habe schon einmal DigiZeitschriften benutzt (Erinnerung: Suche nach dem Aufsatz "Zur Wergeldfrage")

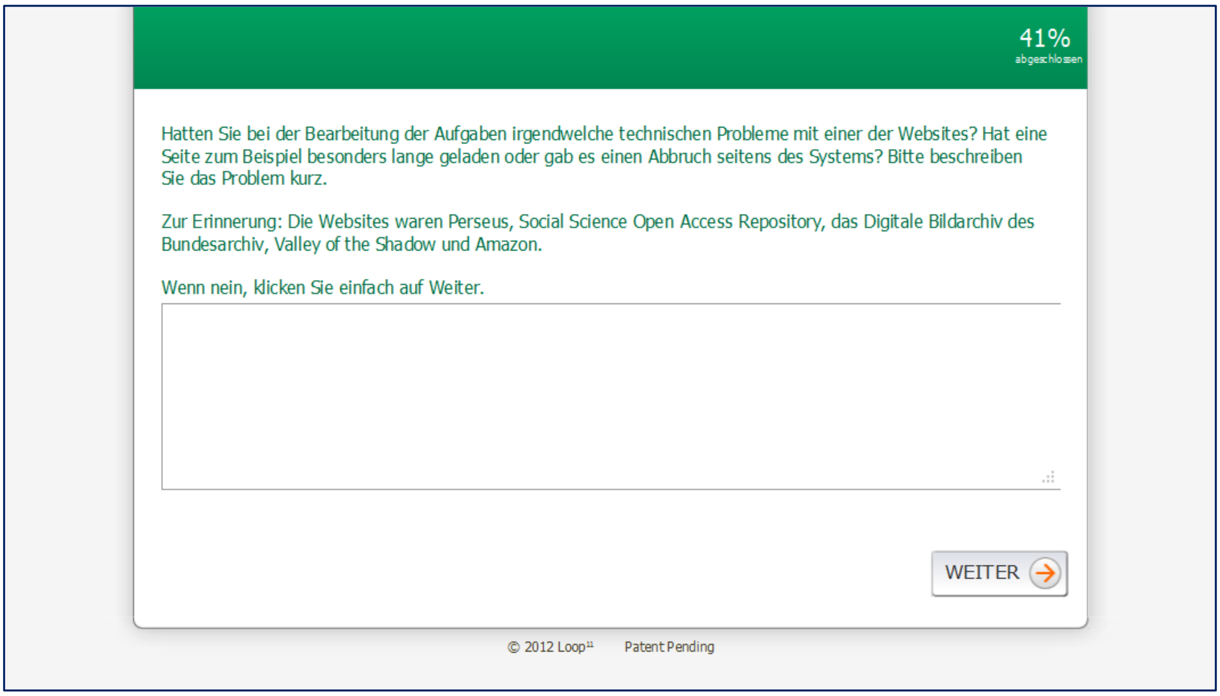
□ Ich habe schon einmal Perseus benutzt (Erinnerung: Suche nach dem Volltext von Antigone)

□ Ich habe schon einmal das Social Science Open Access Repository benutzt (Erinnerung: Suche nach dem Dokument "Wünsche zur Verbesserung der Studiensituation in den Naturwissenschaften")

□ Ich habe schon einmal das Digitale Bildarchiv des Bundesarchivs benutzt (Erinnerung: Suche nach der Fotografie vom CDU Parteitag)

□ Ich habe schon einmal Valley of the Shadow benutzt (Erinnerung: Suche nach dem Brief von Tony Pastor)

Figure 26. Question about prior experiences with one of the digital libraries.

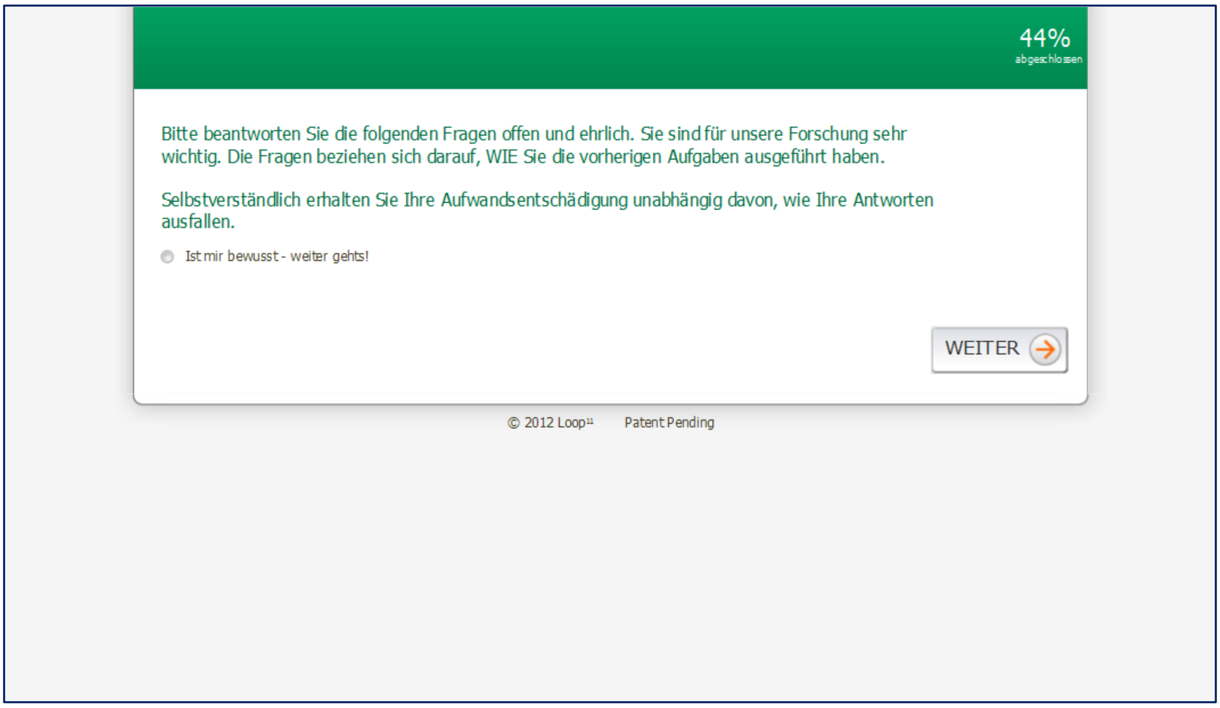


Hatten Sie bei der Bearbeitung der Aufgaben irgendwelche technischen Probleme mit einer der Websites? Hat eine Seite zum Beispiel besonders lange geladen oder gab es einen Abbruch seitens des Systems? Bitte beschreiben Sie das Problem kurz.

Zur Erinnerung: Die Websites waren Perseus, Social Science Open Access Repository, das Digitale Bildarchiv des Bundesarchiv, Valley of the Shadow und Amazon.

Wenn nein, klicken Sie einfach auf Weiter.

Figure 27. Question about potential technical problems during the test.

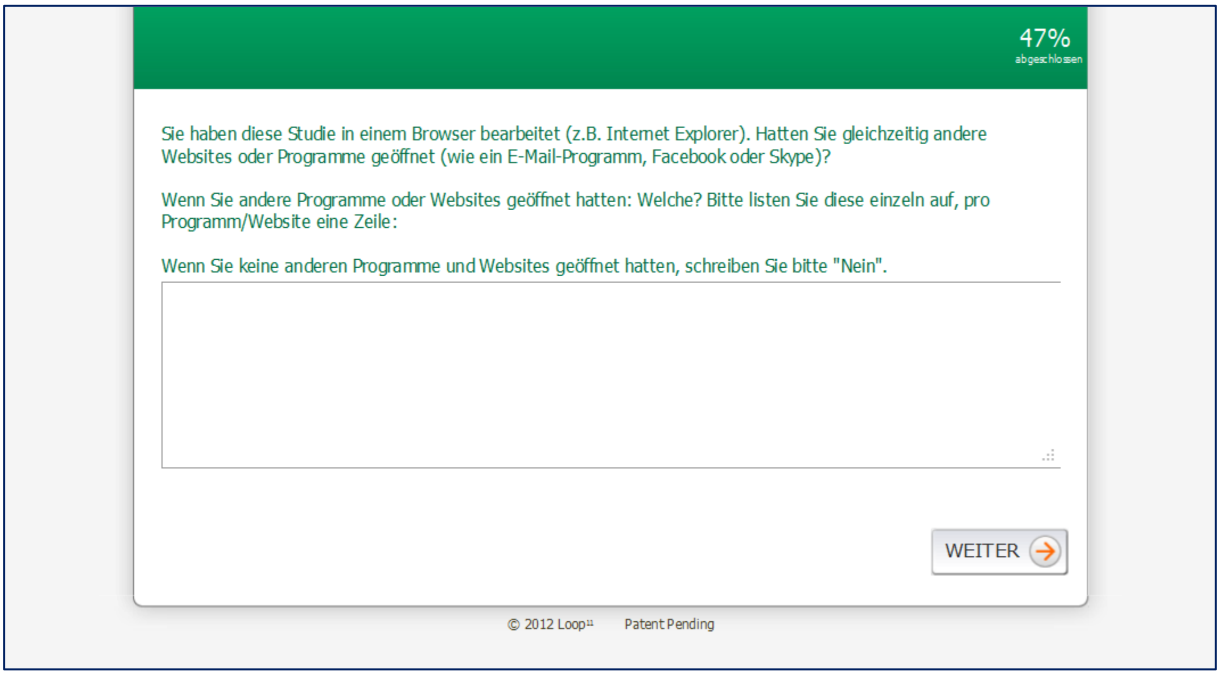


Bitte beantworten Sie die folgenden Fragen offen und ehrlich. Sie sind für unsere Forschung sehr wichtig. Die Fragen beziehen sich darauf, WIE Sie die vorherigen Aufgaben ausgeführt haben.

Selbstverständlich erhalten Sie Ihre Aufwandsentschädigung unabhängig davon, wie Ihre Antworten ausfallen.

○ Ist mir bewusst - weiter gehts!

Figure 28. Break page with assurance that context information has no influence on reward.

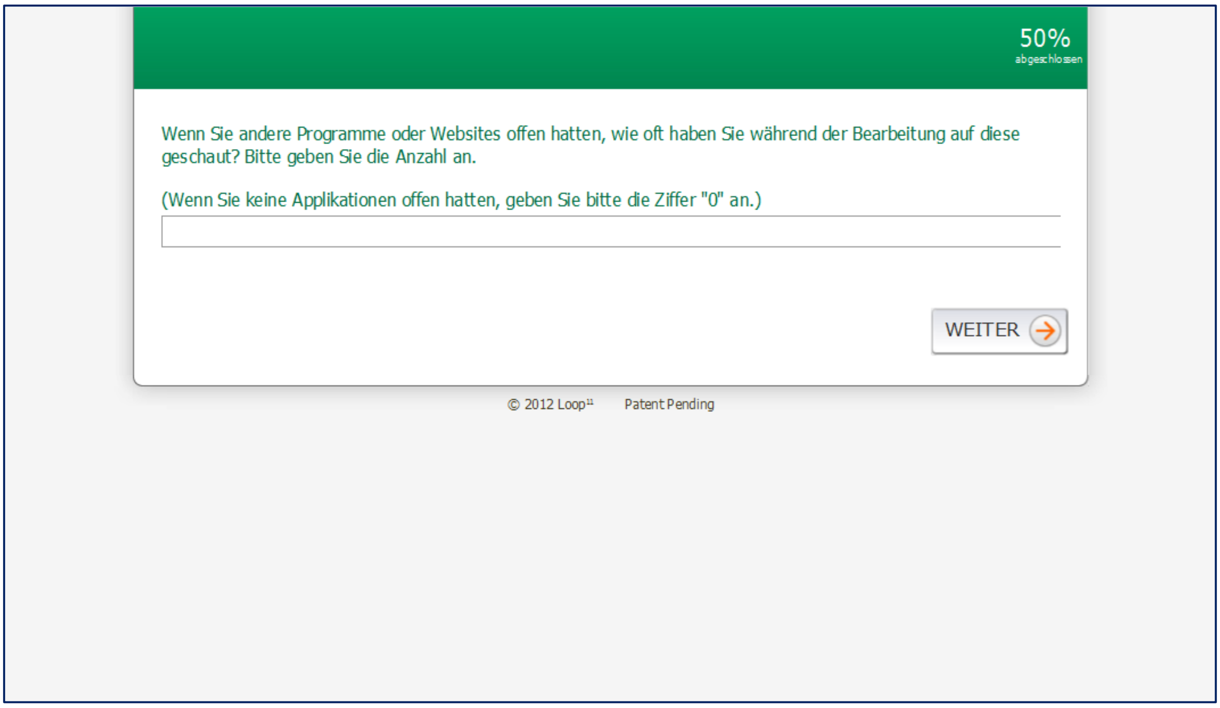


Sie haben diese Studie in einem Browser bearbeitet (z.B. Internet Explorer). Hatten Sie gleichzeitig andere Websites oder Programme geöffnet (wie ein E-Mail-Programm, Facebook oder Skype)?

Wenn Sie andere Programme oder Websites geöffnet hatten: Welche? Bitte listen Sie diese einzeln auf, pro Programm/Website eine Zeile:

Wenn Sie keine anderen Programme und Websites geöffnet hatten, schreiben Sie bitte "Nein".

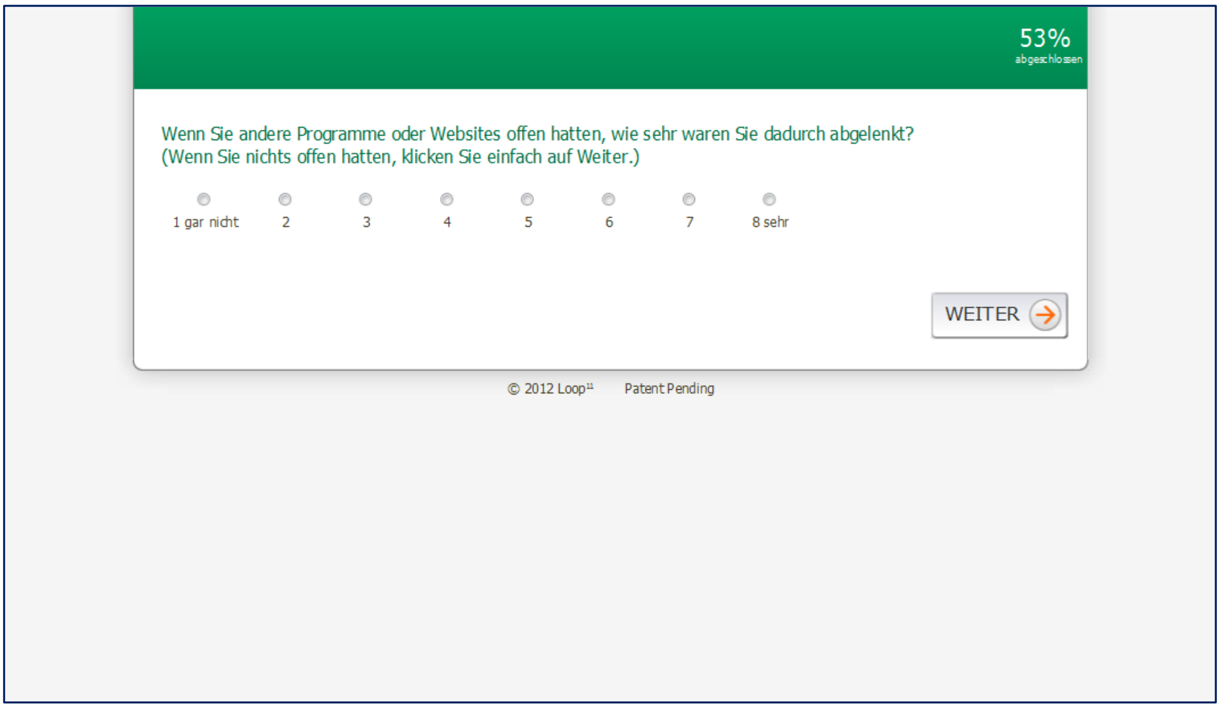
Figure 29. Question about additional open programs.



Wenn Sie andere Programme oder Websites offen hatten, wie oft haben Sie während der Bearbeitung auf diese geschaut? Bitte geben Sie die Anzahl an.

(Wenn Sie keine Applikationen offen hatten, geben Sie bitte die Ziffer "0" an.)

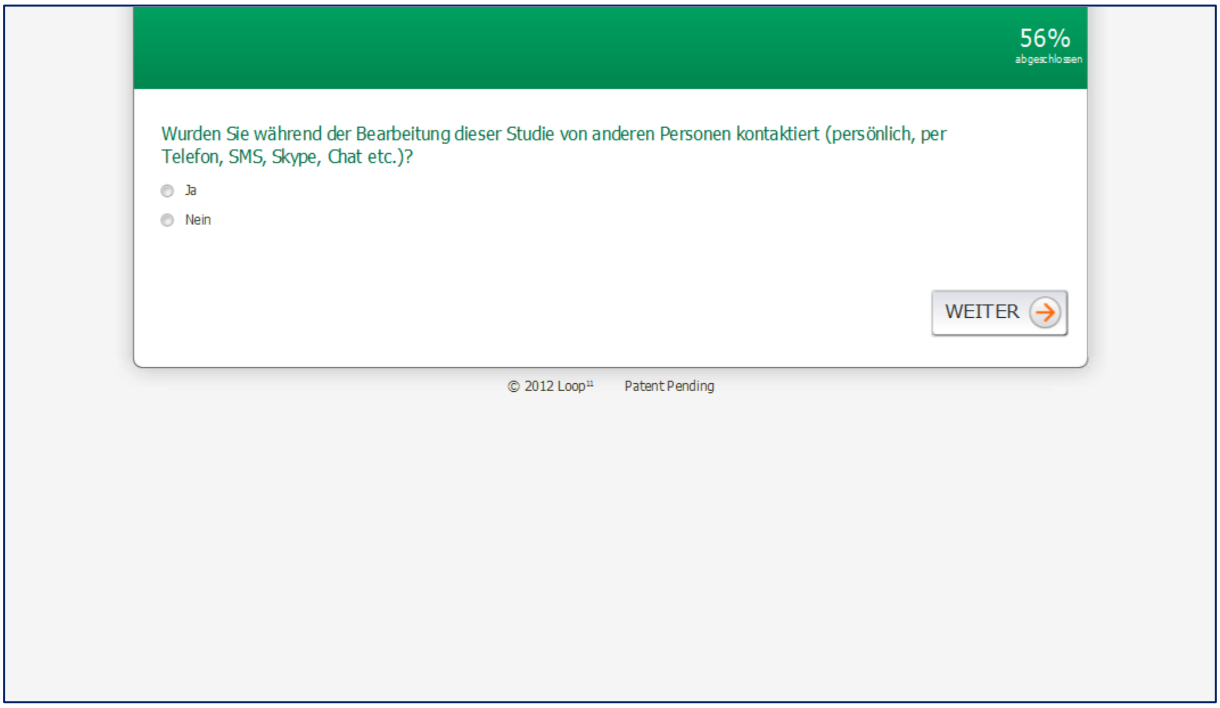
Figure 30. Question about the frequency of looking at the open programs.



Wenn Sie andere Programme oder Websites offen hatten, wie sehr waren Sie dadurch abgelenkt?
(Wenn Sie nichts offen hatten, klicken Sie einfach auf Weiter.)

○ 1 gar nicht ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 sehr

Figure 31. Question about the estimated level of distraction due to open programs.

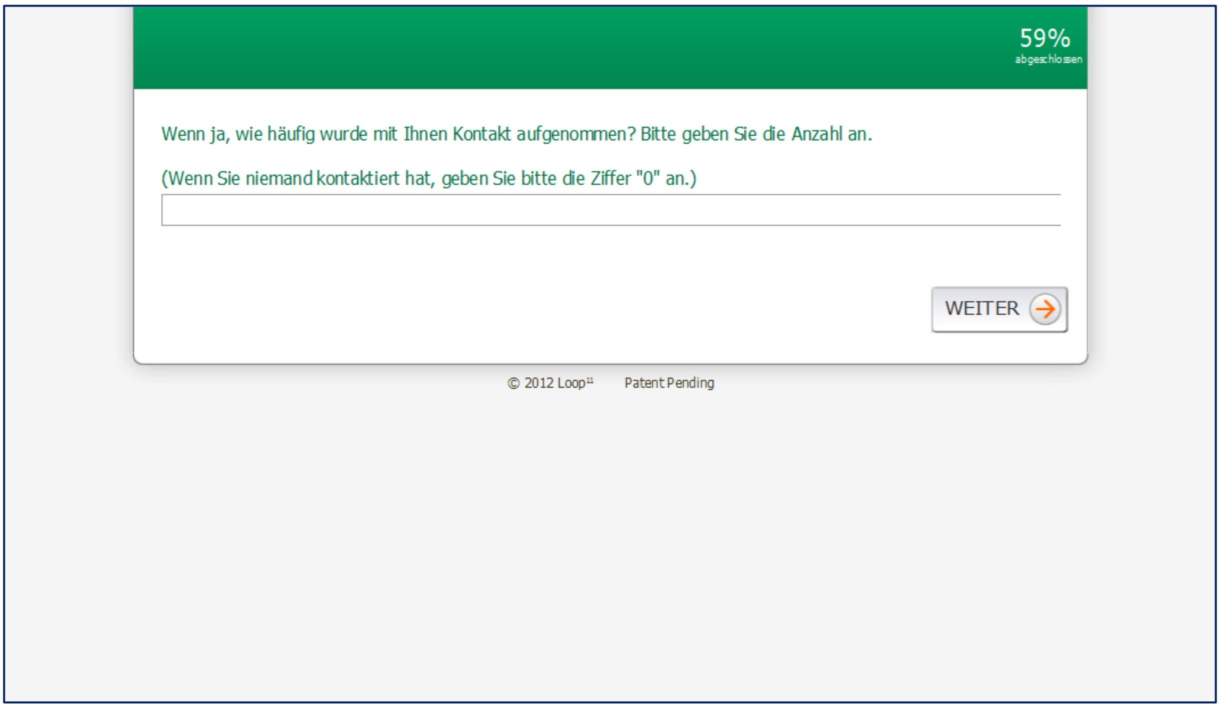


Wurden Sie während der Bearbeitung dieser Studie von anderen Personen kontaktiert (persönlich, per Telefon, SMS, Skype, Chat etc.)?

○ Ja

○ Nein

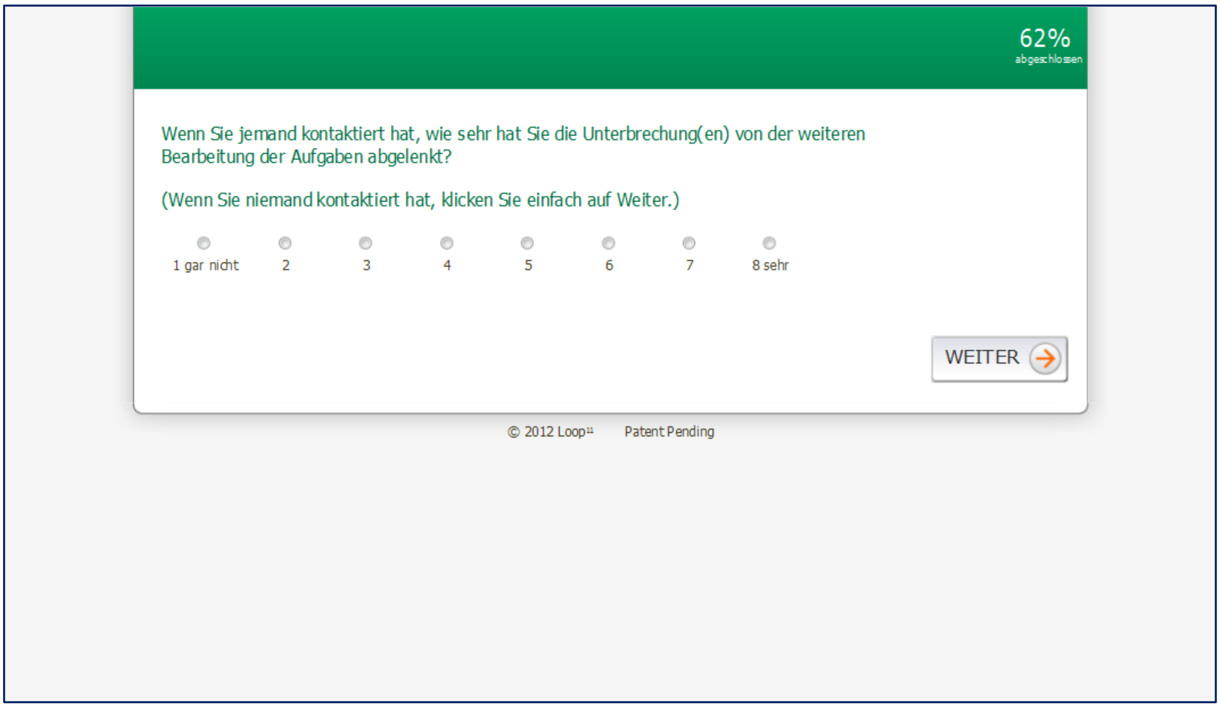
Figure 32. Question about potential contacts during the test.



Wenn ja, wie häufig wurde mit Ihnen Kontakt aufgenommen? Bitte geben Sie die Anzahl an.

(Wenn Sie niemand kontaktiert hat, geben Sie bitte die Ziffer "0" an.)

Figure 33. Question about the frequency of being contacted.

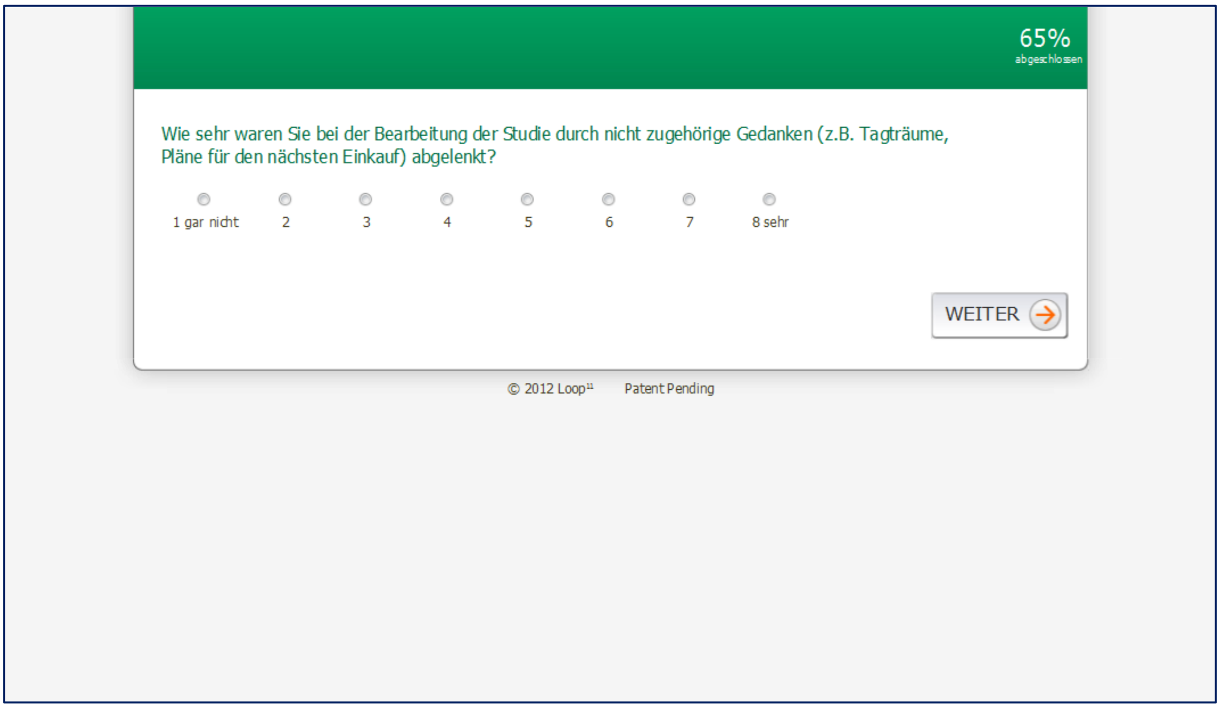


Wenn Sie jemand kontaktiert hat, wie sehr hat Sie die Unterbrechung(en) von der weiteren Bearbeitung der Aufgaben abgelenkt?

(Wenn Sie niemand kontaktiert hat, klicken Sie einfach auf Weiter.)

○ 1 gar nicht ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 sehr

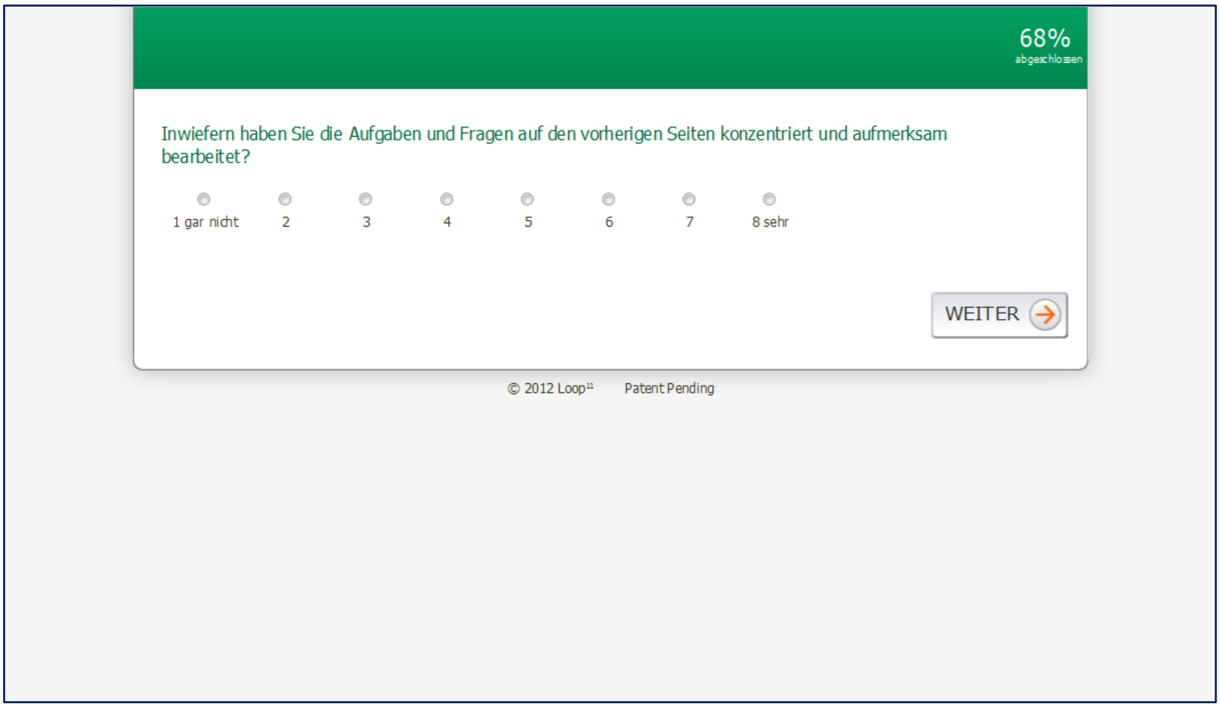
Figure 34. Question about the estimated level of distraction due to contacts.



Wie sehr waren Sie bei der Bearbeitung der Studie durch nicht zugehörige Gedanken (z.B. Tagträume, Pläne für den nächsten Einkauf) abgelenkt?

○ 1 gar nicht ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 sehr

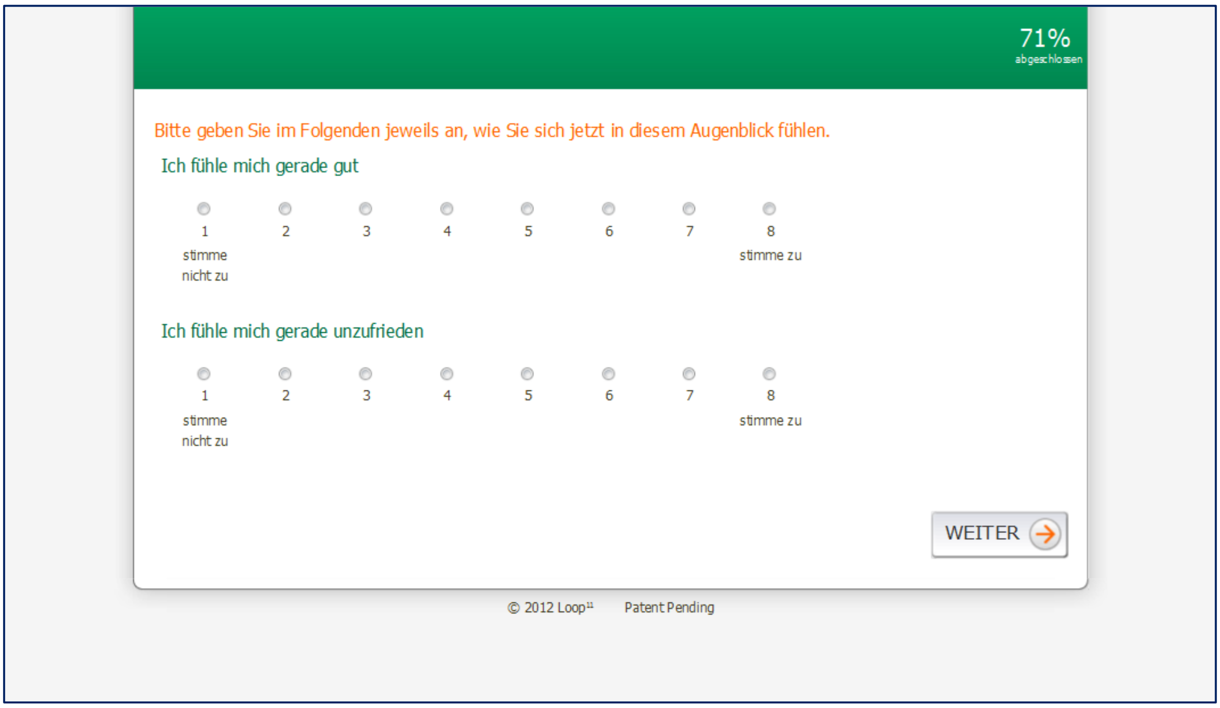
Figure 35. Question about the estimated level of distraction due to daydreams.



Inwiefern haben Sie die Aufgaben und Fragen auf den vorherigen Seiten konzentriert und aufmerksam bearbeitet?

○	○	○	○	○	○	○	○
1 gar nicht	2	3	4	5	6	7	8 sehr

Figure 36. Question about participant's level of attention on the test.



Bitte geben Sie im Folgenden jeweils an, wie Sie sich jetzt in diesem Augenblick fühlen.

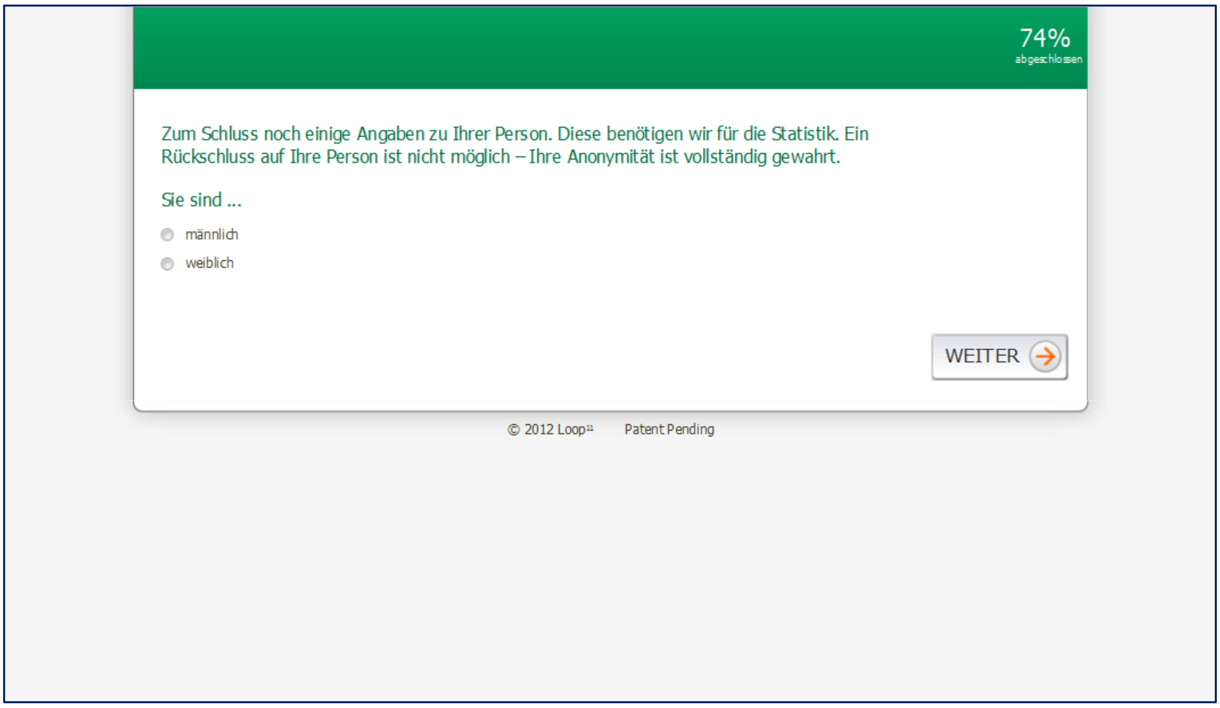
Ich fühle mich gerade gut

○	○	○	○	○	○	○	○
1 stimme nicht zu	2	3	4	5	6	7	8 stimme zu

Ich fühle mich gerade unzufrieden

○	○	○	○	○	○	○	○
1 stimme nicht zu	2	3	4	5	6	7	8 stimme zu

Figure 37. Questions about the participant's current mood.

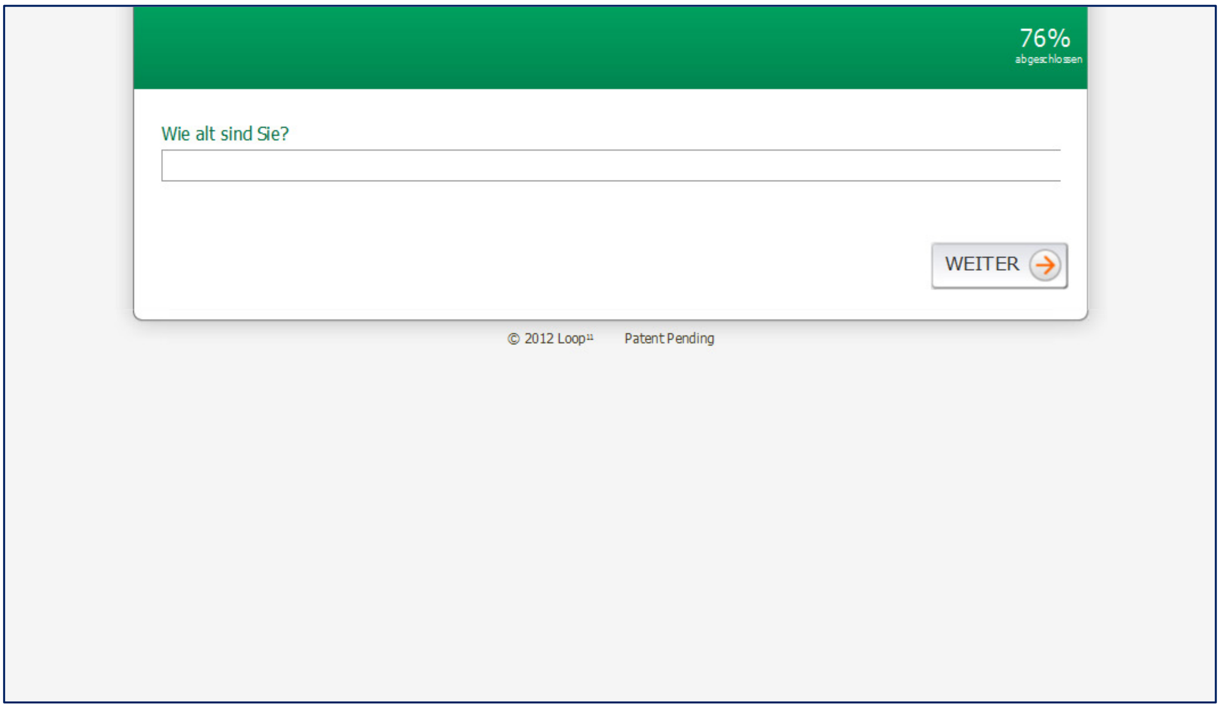


Zum Schluss noch einige Angaben zu Ihrer Person. Diese benötigen wir für die Statistik. Ein Rückschluss auf Ihre Person ist nicht möglich – Ihre Anonymität ist vollständig gewahrt.

Sie sind ...

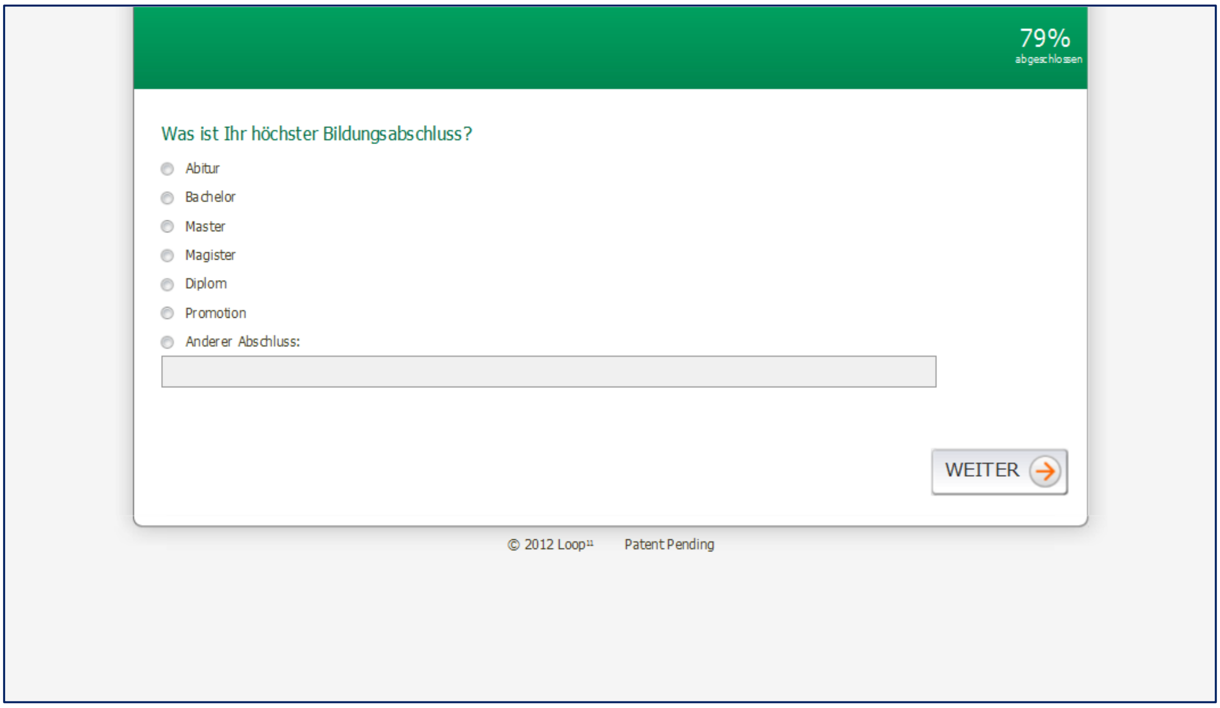
○ männlich
○ weiblich

Figure 38. Question about the participant's gender.



Wie alt sind Sie?

Figure 39. Question about the participant's age.



79%
abgeschlossen

Was ist Ihr höchster Bildungsabschluss?

○ Abitur

○ Bachelor

○ Master

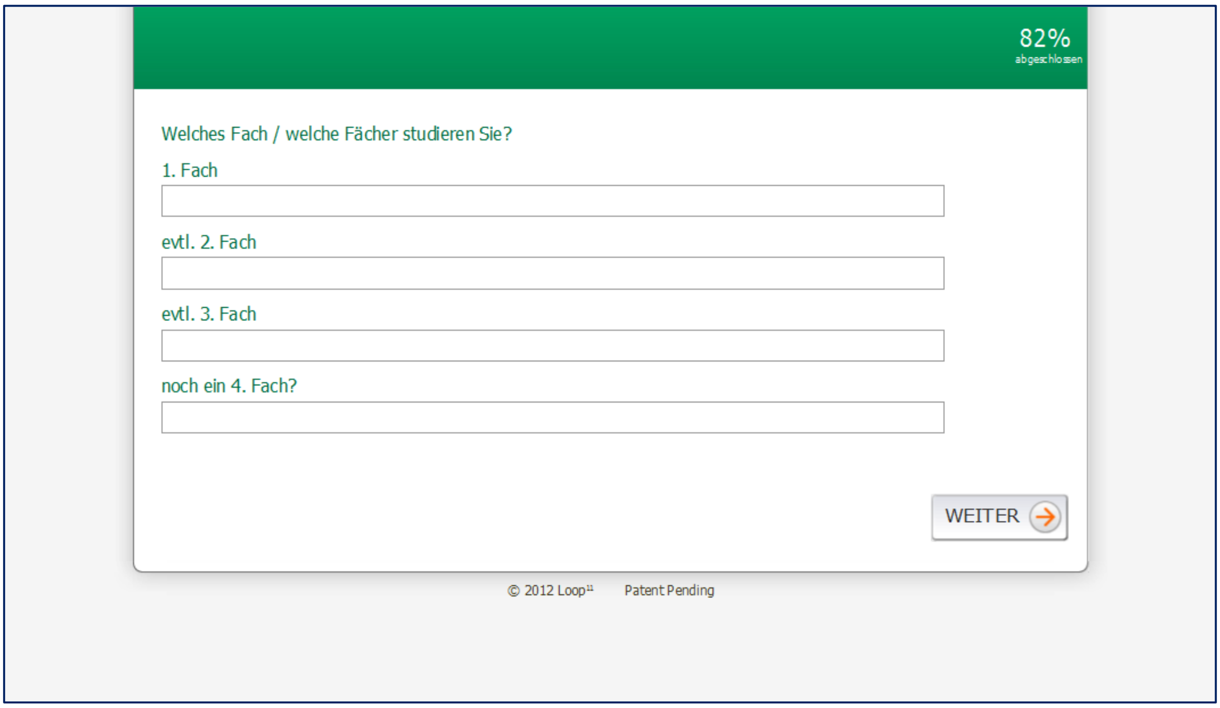
○ Magister

○ Diplom

○ Promotion

○ Anderer Abschluss:

Figure 40. Question about the participant's highest university degree.



82%
abgeschlossen

Welches Fach / welche Fächer studieren Sie?

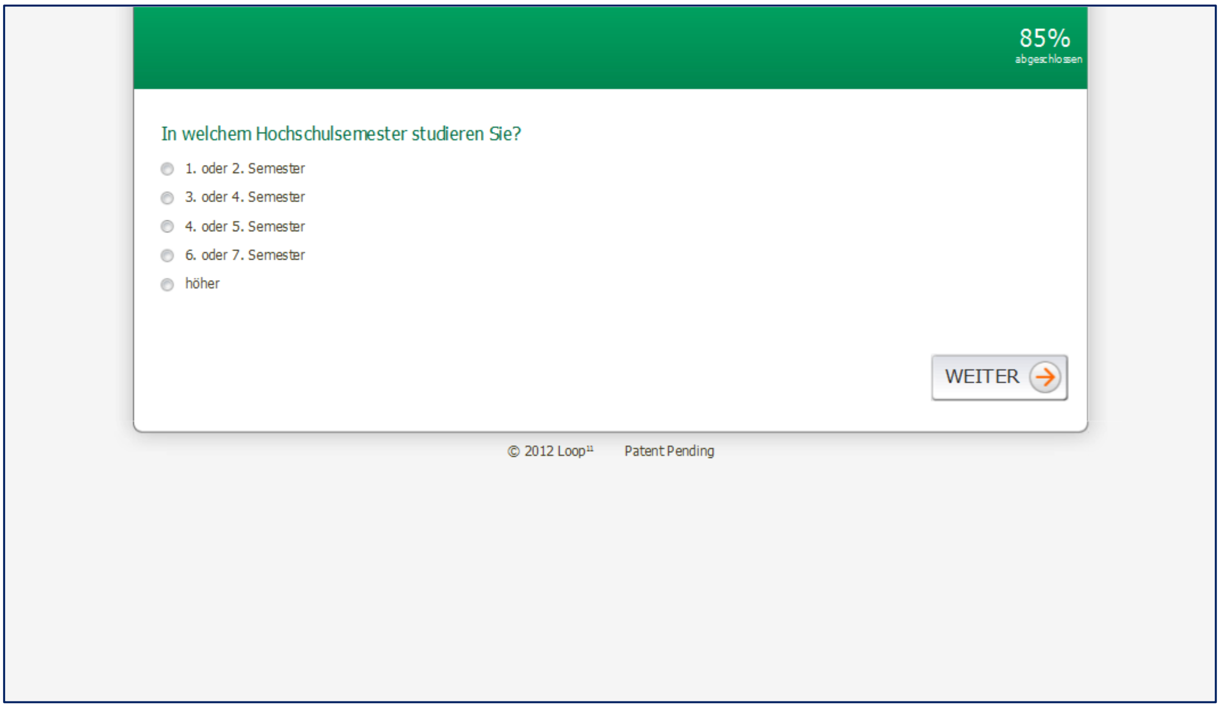
1. Fach

evtl. 2. Fach

evtl. 3. Fach

noch ein 4. Fach?

Figure 41. Question about the participant's first, second or third academic subject.



In welchem Hochschulsemester studieren Sie?

○ 1. oder 2. Semester

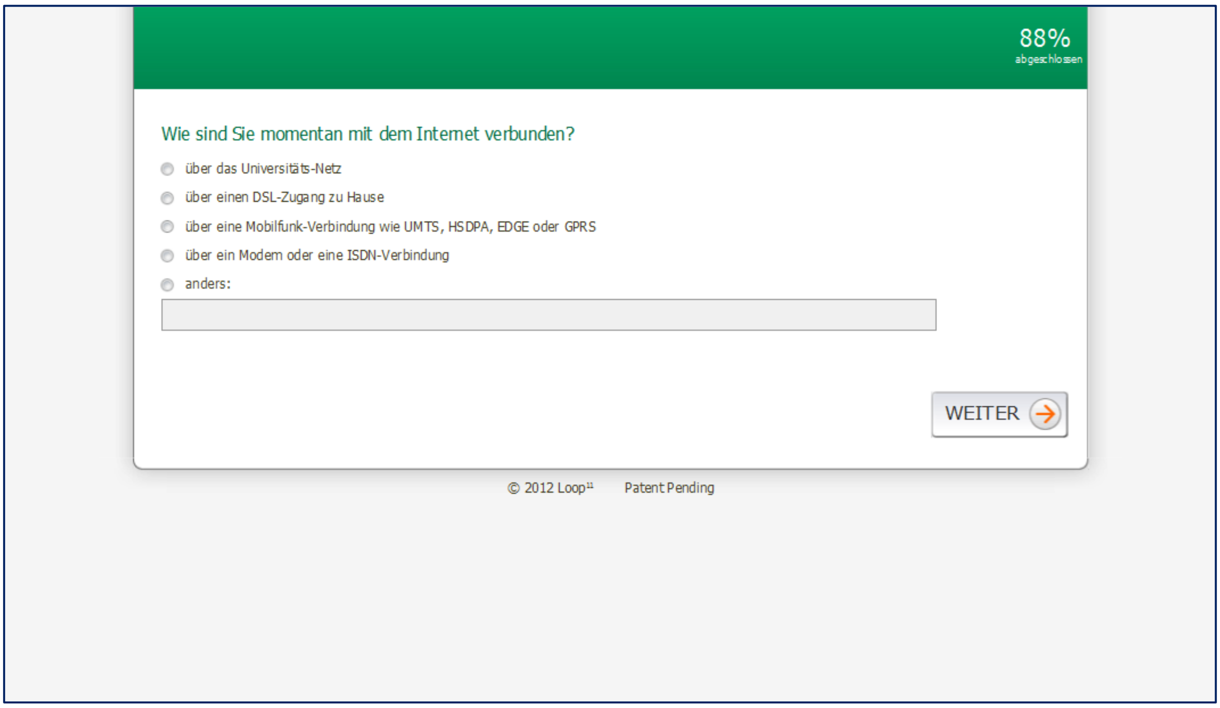
○ 3. oder 4. Semester

○ 4. oder 5. Semester

○ 6. oder 7. Semester

○ höher

Figure 42. Question about the participant's current number of semesters.



Wie sind Sie momentan mit dem Internet verbunden?

○ über das Universitäts-Netz

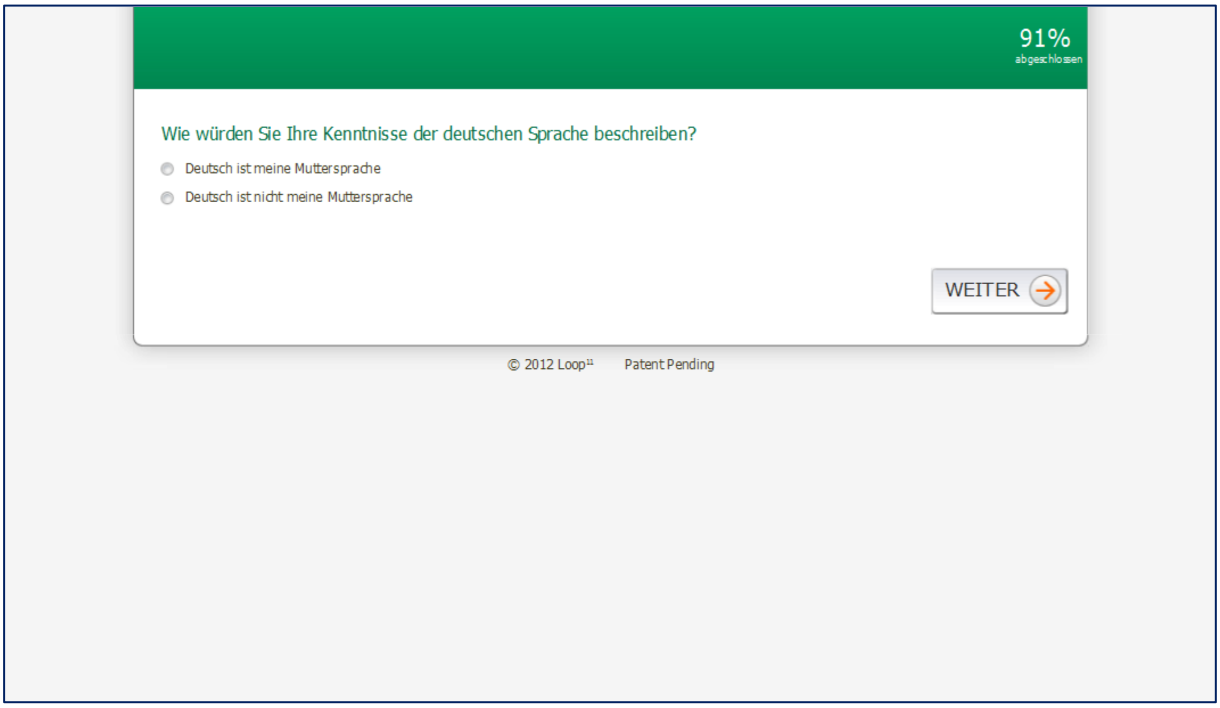
○ über einen DSL-Zugang zu Hause

○ über eine Mobilfunk-Verbindung wie UMTS, HSDPA, EDGE oder GPRS

○ über ein Modem oder eine ISDN-Verbindung

○ anders:

Figure 43. Question about participant's kind of internet connection.



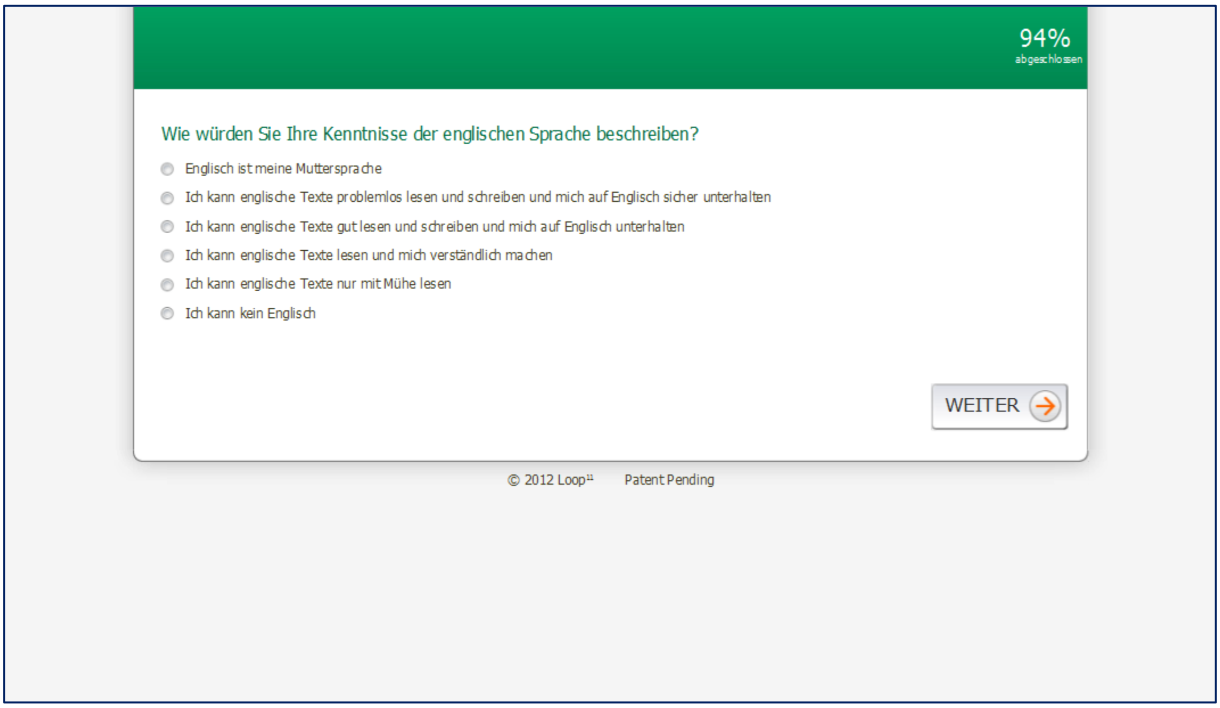
Wie würden Sie Ihre Kenntnisse der deutschen Sprache beschreiben?

○ Deutsch ist meine Muttersprache

○ Deutsch ist nicht meine Muttersprache

© 2012 Loop11 Patent Pending

Figure 44. Question about the participant's German language skills.



Wie würden Sie Ihre Kenntnisse der englischen Sprache beschreiben?

○ Englisch ist meine Muttersprache

○ Ich kann englische Texte problemlos lesen und schreiben und mich auf Englisch sicher unterhalten

○ Ich kann englische Texte gut lesen und schreiben und mich auf Englisch unterhalten

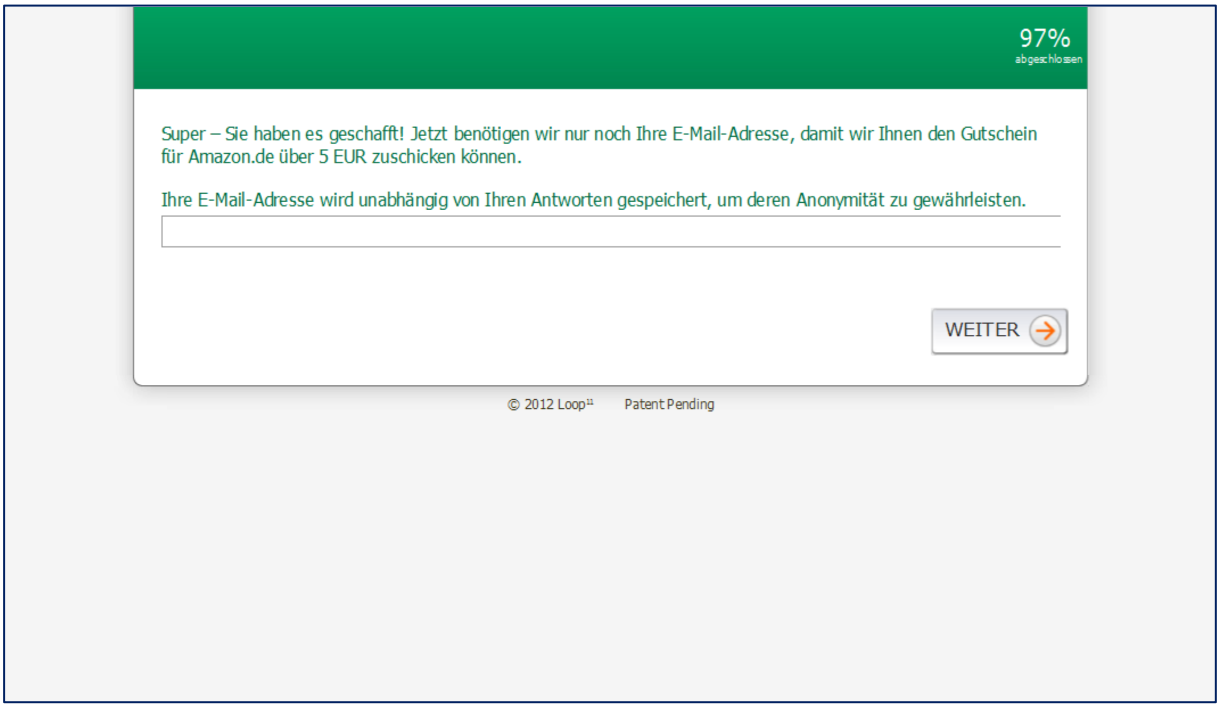
○ Ich kann englische Texte lesen und mich verständlich machen

○ Ich kann englische Texte nur mit Mühe lesen

○ Ich kann kein Englisch

© 2012 Loop11 Patent Pending

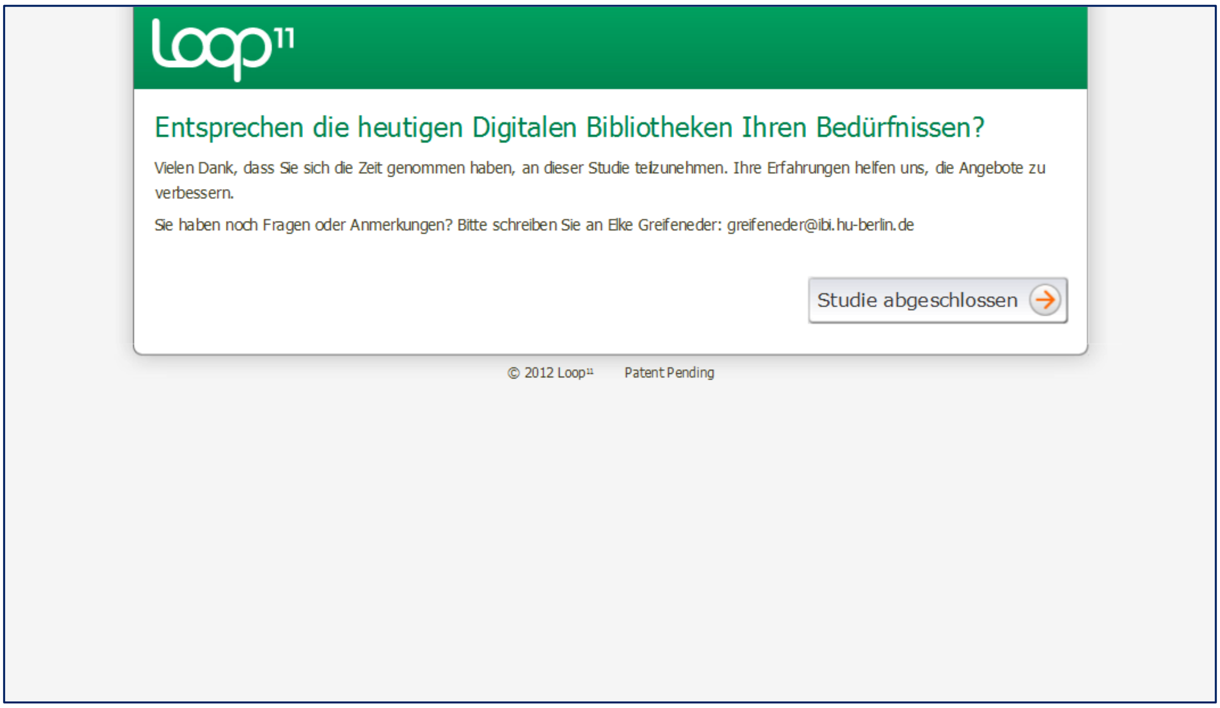
Figure 45. Question about the participant's English language skills.



Super – Sie haben es geschafft! Jetzt benötigen wir nur noch Ihre E-Mail-Adresse, damit wir Ihnen den Gutschein für Amazon.de über 5 EUR zuschicken können.

Ihre E-Mail-Adresse wird unabhängig von Ihren Antworten gespeichert, um deren Anonymität zu gewährleisten.

Figure 46. Final page asking for an email address for the reward (not required).



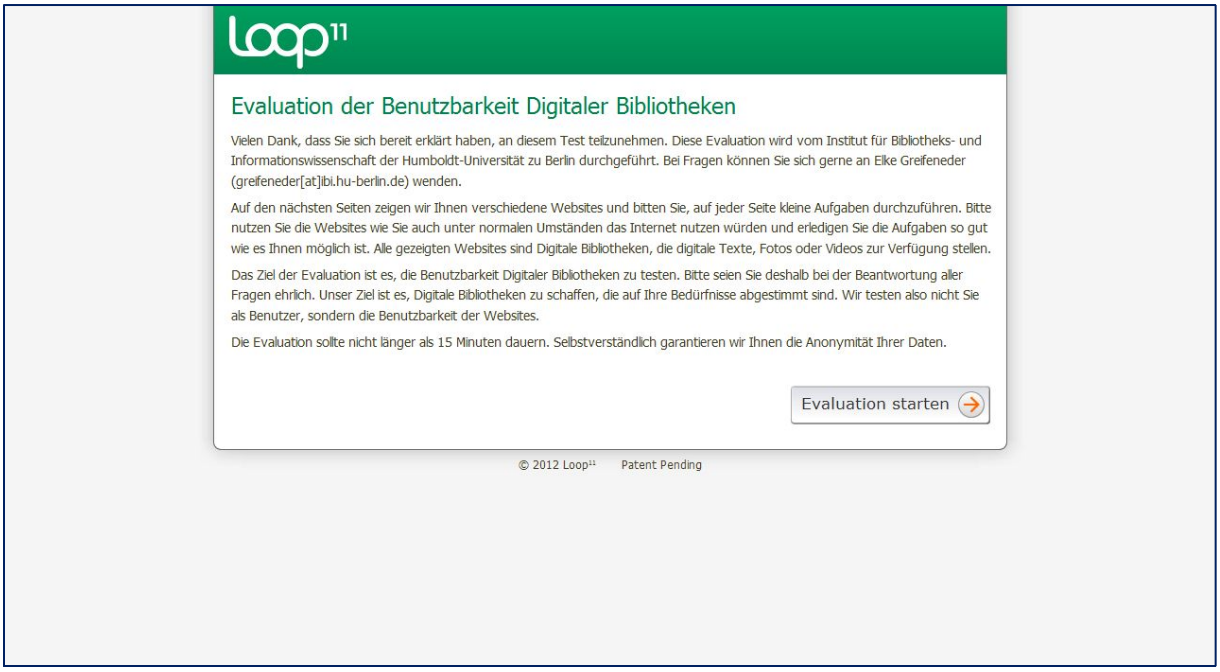
Entsprechen die heutigen Digitalen Bibliotheken Ihren Bedürfnissen?

Vielen Dank, dass Sie sich die Zeit genommen haben, an dieser Studie teilzunehmen. Ihre Erfahrungen helfen uns, die Angebote zu verbessern.

Sie haben noch Fragen oder Anmerkungen? Bitte schreiben Sie an Elke Greifeneder: greifeneder@ibi.hu-berlin.de

Figure 47. Good-bye page with contact details.

Appendix 4: Screenshots of the pilot test



Evaluation der Benutzbarkeit Digitaler Bibliotheken

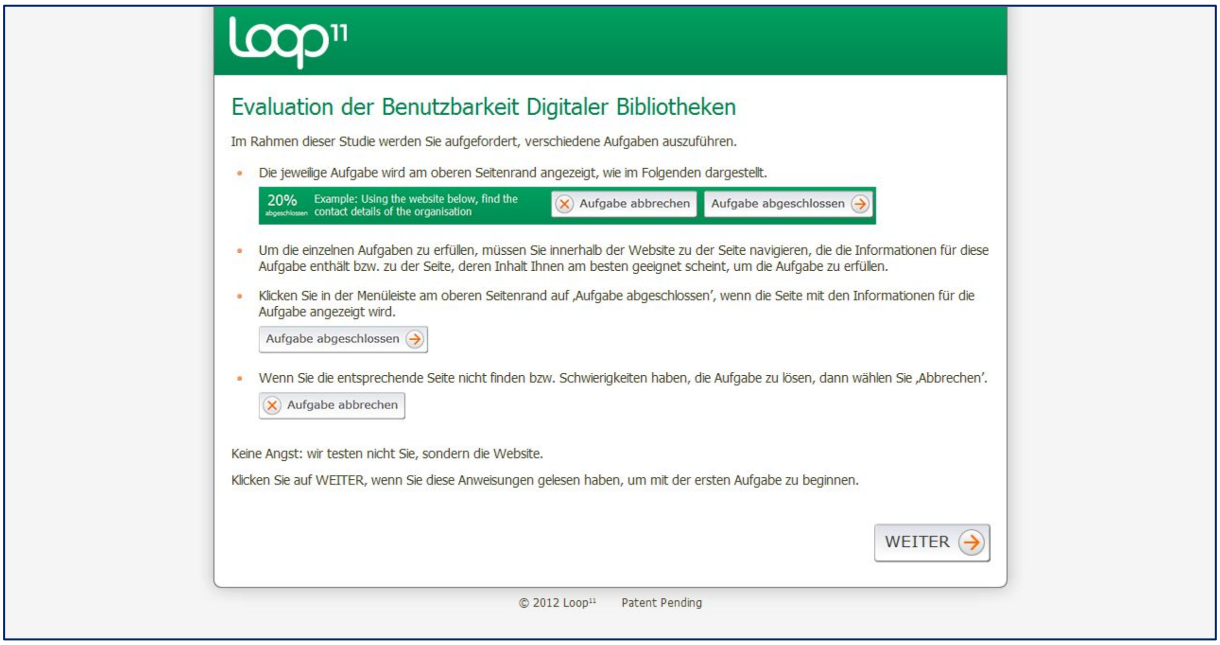
Vielen Dank, dass Sie sich bereit erklärt haben, an diesem Test teilzunehmen. Diese Evaluation wird vom Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin durchgeführt. Bei Fragen können Sie sich gerne an Elke Greifeneder (greifeneder[at]ibi.hu-berlin.de) wenden.

Auf den nächsten Seiten zeigen wir Ihnen verschiedene Websites und bitten Sie, auf jeder Seite kleine Aufgaben durchzuführen. Bitte nutzen Sie die Websites wie Sie auch unter normalen Umständen das Internet nutzen würden und erledigen Sie die Aufgaben so gut wie es Ihnen möglich ist. Alle gezeigten Websites sind Digitale Bibliotheken, die digitale Texte, Fotos oder Videos zur Verfügung stellen.

Das Ziel der Evaluation ist es, die Benutzbarkeit Digitaler Bibliotheken zu testen. Bitte seien Sie deshalb bei der Beantwortung aller Fragen ehrlich. Unser Ziel ist es, Digitale Bibliotheken zu schaffen, die auf Ihre Bedürfnisse abgestimmt sind. Wir testen also nicht Sie als Benutzer, sondern die Benutzbarkeit der Websites.

Die Evaluation sollte nicht länger als 15 Minuten dauern. Selbstverständlich garantieren wir Ihnen die Anonymität Ihrer Daten.

Figure 48. Pilot test: Welcome page.



Evaluation der Benutzbarkeit Digitaler Bibliotheken

Im Rahmen dieser Studie werden Sie aufgefordert, verschiedene Aufgaben auszuführen.

- Die jeweilige Aufgabe wird am oberen Seitenrand angezeigt, wie im Folgenden dargestellt.

20% abgeschlossen Example: Using the website below, find the contact details of the organisation

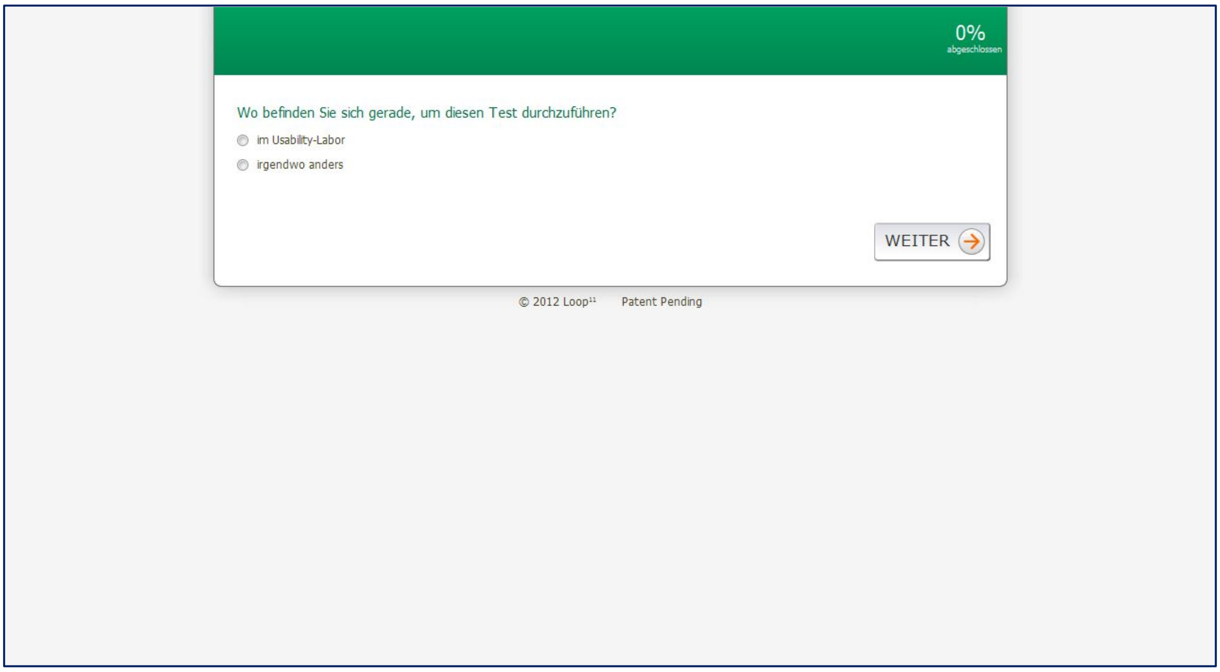
- Um die einzelnen Aufgaben zu erfüllen, müssen Sie innerhalb der Website zu der Seite navigieren, die die Informationen für diese Aufgabe enthält bzw. zu der Seite, deren Inhalt Ihnen am besten geeignet scheint, um die Aufgabe zu erfüllen.
- Klicken Sie in der Menüleiste am oberen Seitenrand auf ‚Aufgabe abgeschlossen', wenn die Seite mit den Informationen für die Aufgabe angezeigt wird.

- Wenn Sie die entsprechende Seite nicht finden bzw. Schwierigkeiten haben, die Aufgabe zu lösen, dann wählen Sie ‚Abbrechen'.

Keine Angst: wir testen nicht Sie, sondern die Website.

Klicken Sie auf WEITER, wenn Sie diese Anweisungen gelesen haben, um mit der ersten Aufgabe zu beginnen.

Figure 49. Pilot test: Help page about test functionality.

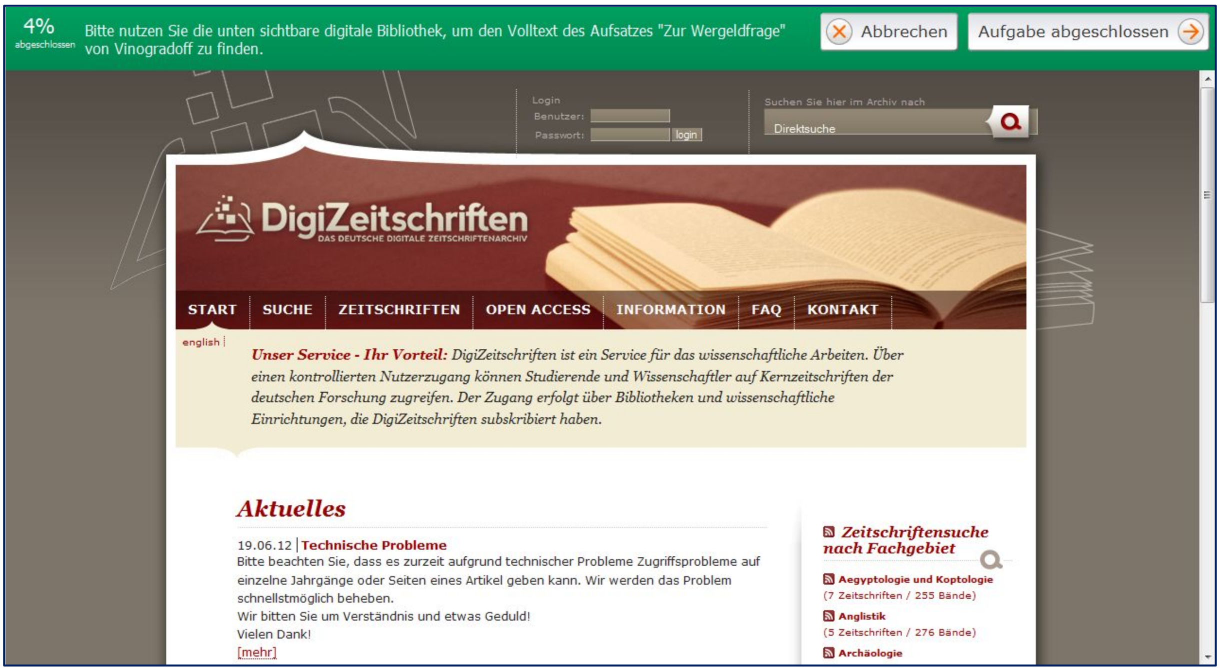


Wo befinden Sie sich gerade, um diesen Test durchzuführen?

○ im Usability-Labor

○ irgendwo anders

Figure 50. Pilot test: Question about the setting.



Bitte nutzen Sie die unten sichtbare digitale Bibliothek, um den Volltext des Aufsatzes "Zur Wergeldfrage" von Vinogradoff zu finden.

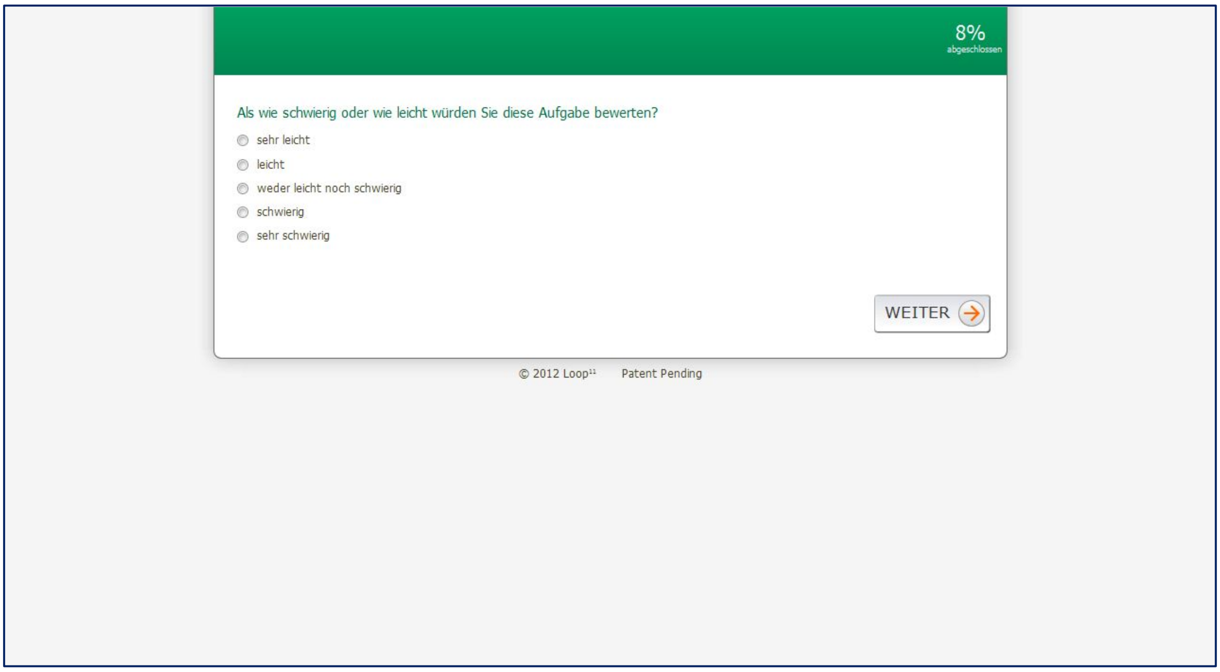
Unser Service - Ihr Vorteil: DigiZeitschriften ist ein Service für das wissenschaftliche Arbeiten. Über einen kontrollierten Nutzerzugang können Studierende und Wissenschaftler auf Kernzeitschriften der deutschen Forschung zugreifen. Der Zugang erfolgt über Bibliotheken und wissenschaftliche Einrichtungen, die DigiZeitschriften subskribiert haben.

Aktuelles

19.06.12 | **Technische Probleme**
Bitte beachten Sie, dass es zurzeit aufgrund technischer Probleme Zugriffsprobleme auf einzelne Jahrgänge oder Seiten eines Artikel geben kann. Wir werden das Problem schnellstmöglich beheben.
Wir bitten Sie um Verständnis und etwas Geduld!
Vielen Dank!
[mehr]

Zeitschriftensuche nach Fachgebiet
Aegyptologie und Koptologie
(7 Zeitschriften / 255 Bände)
Anglistik
(5 Zeitschriften / 276 Bände)
Archäologie

Figure 51. Pilot test: First task in *DigiZeitschriften*.

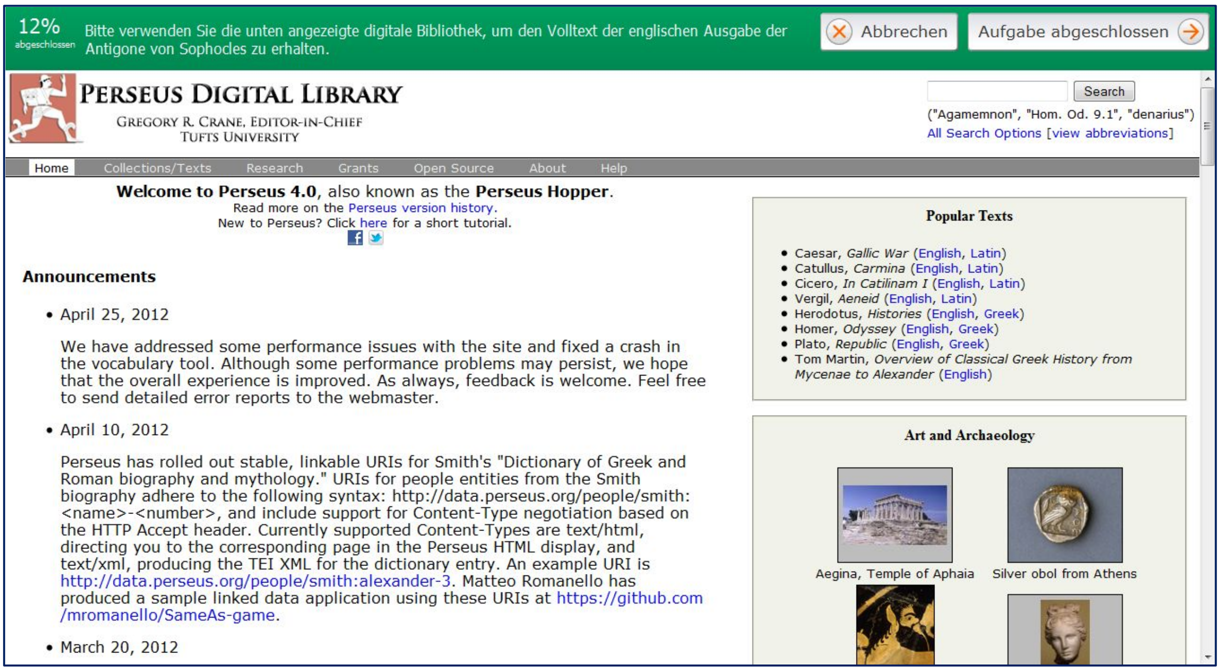


Als wie schwierig oder wie leicht würden Sie diese Aufgabe bewerten?

○ sehr leicht
○ leicht
○ weder leicht noch schwierig
○ schwierig
○ sehr schwierig

© 2012 Loop11 Patent Pending

Figure 52. Pilot test: Question about task difficulty after each task.



Bitte verwenden Sie die unten angezeigte digitale Bibliothek, um den Volltext der englischen Ausgabe der Antigone von Sophocles zu erhalten.

Welcome to Perseus 4.0, also known as the **Perseus Hopper**.
Read more on the Perseus version history.
New to Perseus? Click here for a short tutorial.

Announcements

- April 25, 2012

  We have addressed some performance issues with the site and fixed a crash in the vocabulary tool. Although some performance problems may persist, we hope that the overall experience is improved. As always, feedback is welcome. Feel free to send detailed error reports to the webmaster.

- April 10, 2012

  Perseus has rolled out stable, linkable URIs for Smith's "Dictionary of Greek and Roman biography and mythology." URIs for people entities from the Smith biography adhere to the following syntax: http://data.perseus.org/people/smith:<name>-<number>, and include support for Content-Type negotiation based on the HTTP Accept header. Currently supported Content-Types are text/html, directing you to the corresponding page in the Perseus HTML display, and text/xml, producing the TEI XML for the dictionary entry. An example URI is http://data.perseus.org/people/smith:alexander-3. Matteo Romanello has produced a sample linked data application using these URIs at https://github.com/mromanello/SameAs-game.

- March 20, 2012

Popular Texts

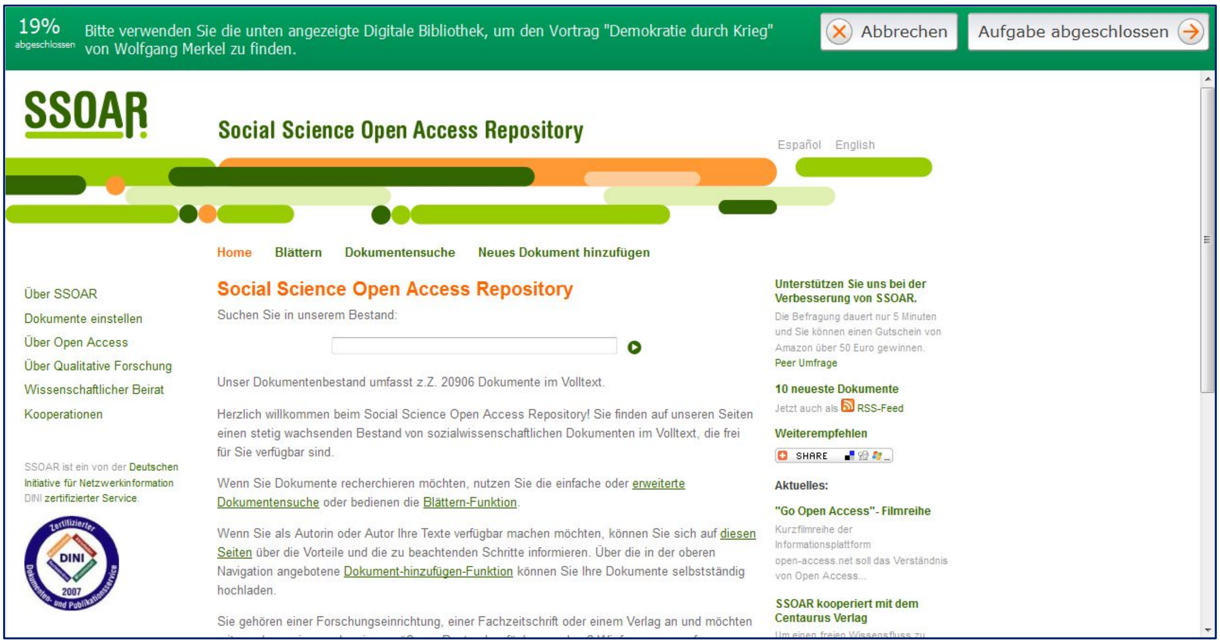
- Caesar, *Gallic War* (English, Latin)
- Catullus, *Carmina* (English, Latin)
- Cicero, *In Catilinam I* (English, Latin)
- Vergil, *Aeneid* (English, Latin)
- Herodotus, *Histories* (English, Greek)
- Homer, *Odyssey* (English, Greek)
- Plato, *Republic* (English, Greek)
- Tom Martin, *Overview of Classical Greek History from Mycenae to Alexander* (English)

Art and Archaeology

Aegina, Temple of Aphaia

Silver obol from Athens

Figure 53. Pilot test: Second task in Perseus.



Bitte verwenden Sie die unten angezeigte Digitale Bibliothek, um den Vortrag "Demokratie durch Krieg" von Wolfgang Merkel zu finden.

Social Science Open Access Repository

Social Science Open Access Repository

Suchen Sie in unserem Bestand:

Unser Dokumentenbestand umfasst z.Z. 20906 Dokumente im Volltext.

Herzlich willkommen beim Social Science Open Access Repository! Sie finden auf unseren Seiten einen stetig wachsenden Bestand von sozialwissenschaftlichen Dokumenten im Volltext, die frei für Sie verfügbar sind.

Wenn Sie Dokumente recherchieren möchten, nutzen Sie die einfache oder erweiterte Dokumentensuche oder bedienen die Blättern-Funktion.

Wenn Sie als Autorin oder Autor Ihre Texte verfügbar machen möchten, können Sie sich auf diesen Seiten über die Vorteile und die zu beachtenden Schritte informieren. Über die in der oberen Navigation angebotene Dokument-hinzufügen-Funktion können Sie Ihre Dokumente selbstständig hochladen.

Sie gehören einer Forschungseinrichtung, einer Fachzeitschrift oder einem Verlag an und möchten

SSOAR ist ein von der Deutschen Initiative für Netzwerkinformation DINI zertifizierter Service.

Unterstützen Sie uns bei der Verbesserung von SSOAR.

Die Befragung dauert nur 5 Minuten und Sie können einen Gutschein von Amazon über 50 Euro gewinnen.

Peer Umfrage

10 neueste Dokumente

Jetzt auch als RSS-Feed

Weiterempfehlen

Aktuelles:

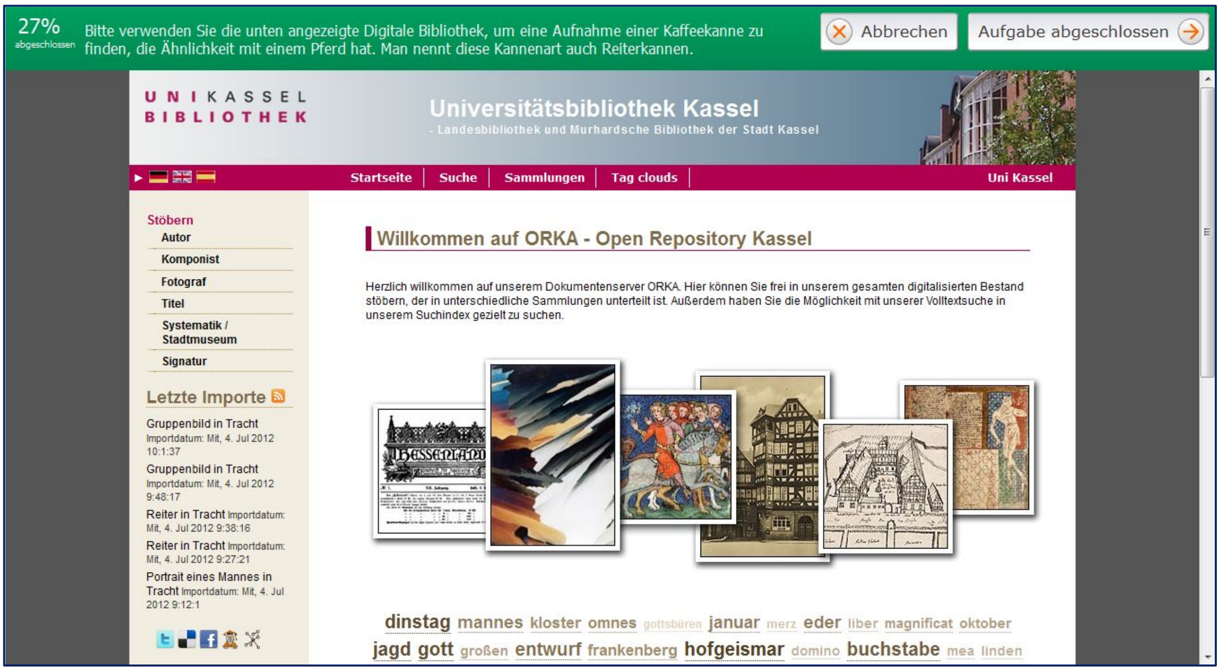
"Go Open Access"- Filmreihe

Kurzfilmreihe der Informationsplattform open-access.net soll das Verständnis von Open Access...

SSOAR kooperiert mit dem Centaurus Verlag

Um einen freien Wissensfluss zu

Figure 54. Pilot test: Third task in SSOAR.



Bitte verwenden Sie die unten angezeigte Digitale Bibliothek, um eine Aufnahme einer Kaffeekanne zu finden, die Ähnlichkeit mit einem Pferd hat. Man nennt diese Kannenart auch Reiterkannen.

Universitätsbibliothek Kassel

- Landesbibliothek und Murhardsche Bibliothek der Stadt Kassel

Willkommen auf ORKA - Open Repository Kassel

Herzlich willkommen auf unserem Dokumentenserver ORKA. Hier können Sie frei in unserem gesamten digitalisierten Bestand stöbern, der in unterschiedliche Sammlungen unterteilt ist. Außerdem haben Sie die Möglichkeit mit unserer Volltextsuche in unserem Suchindex gezielt zu suchen.

Letzte Importe

Gruppenbild in Tracht Importdatum: Mit, 4. Jul 2012 10:1:37

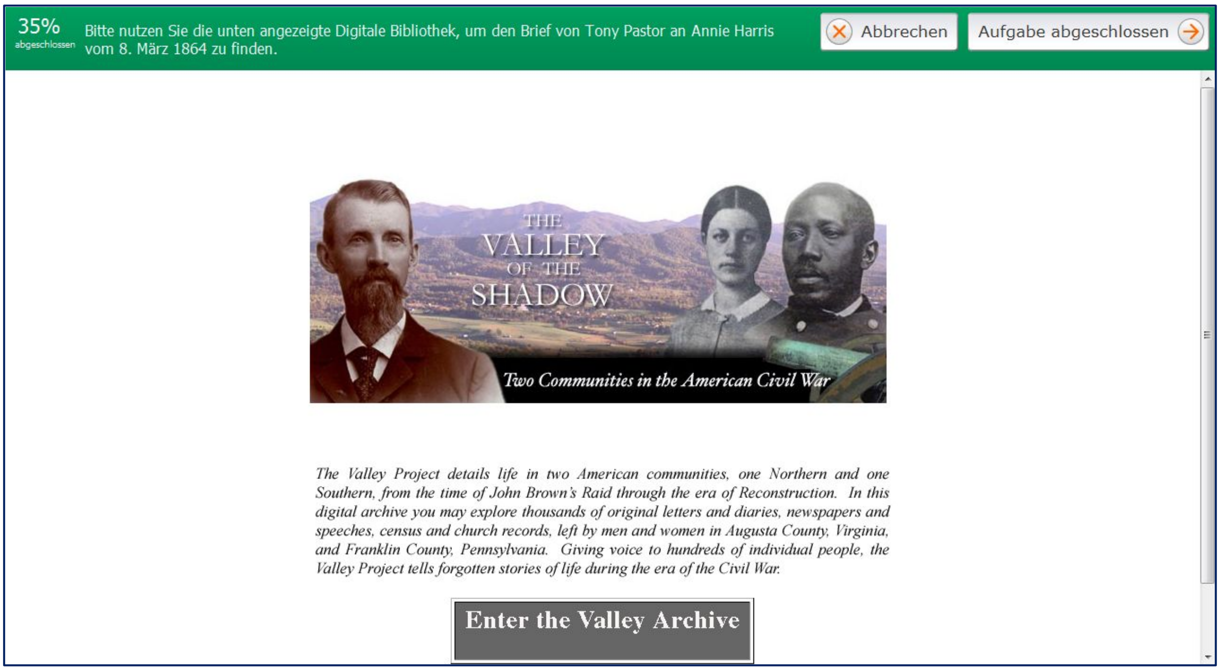
Gruppenbild in Tracht Importdatum: Mit, 4. Jul 2012 9:48:17

Reiter in Tracht Importdatum: Mit, 4. Jul 2012 9:38:16

Reiter in Tracht Importdatum: Mit, 4. Jul 2012 9:27:21

Portrait eines Mannes in Tracht Importdatum: Mit, 4. Jul 2012 9:12:1

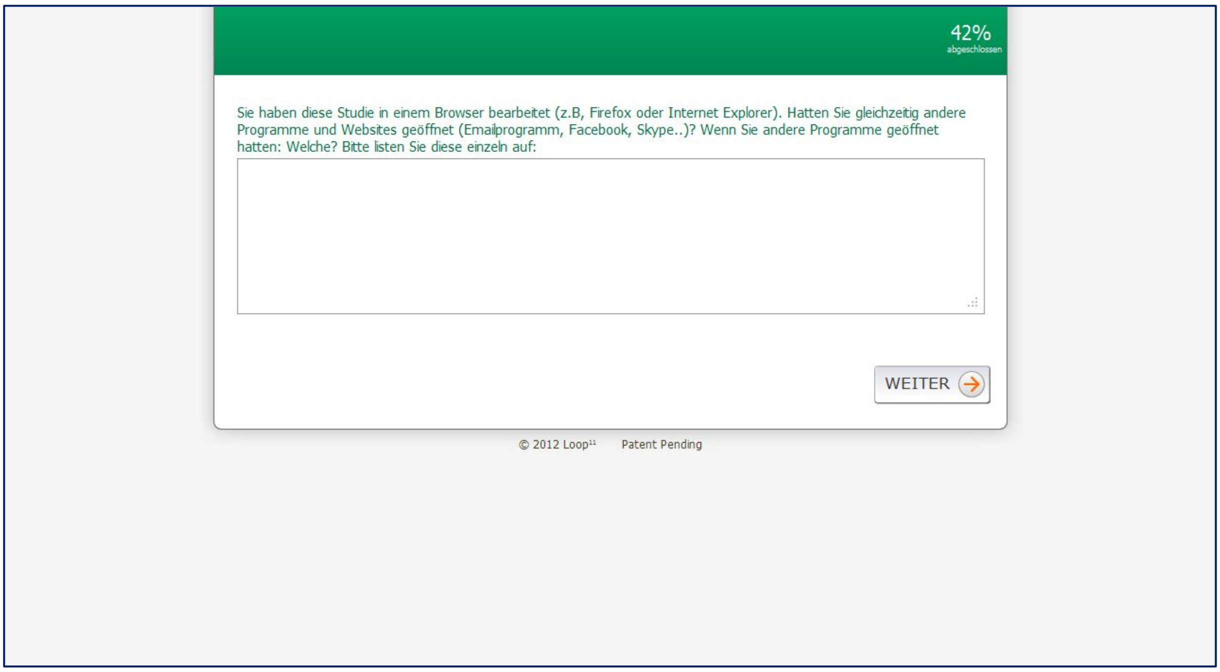
Figure 55. Pilot test: Fourth task in ORKA.



35% abgeschlossen Bitte nutzen Sie die unten angezeigte Digitale Bibliothek, um den Brief von Tony Pastor an Annie Harris vom 8. März 1864 zu finden.

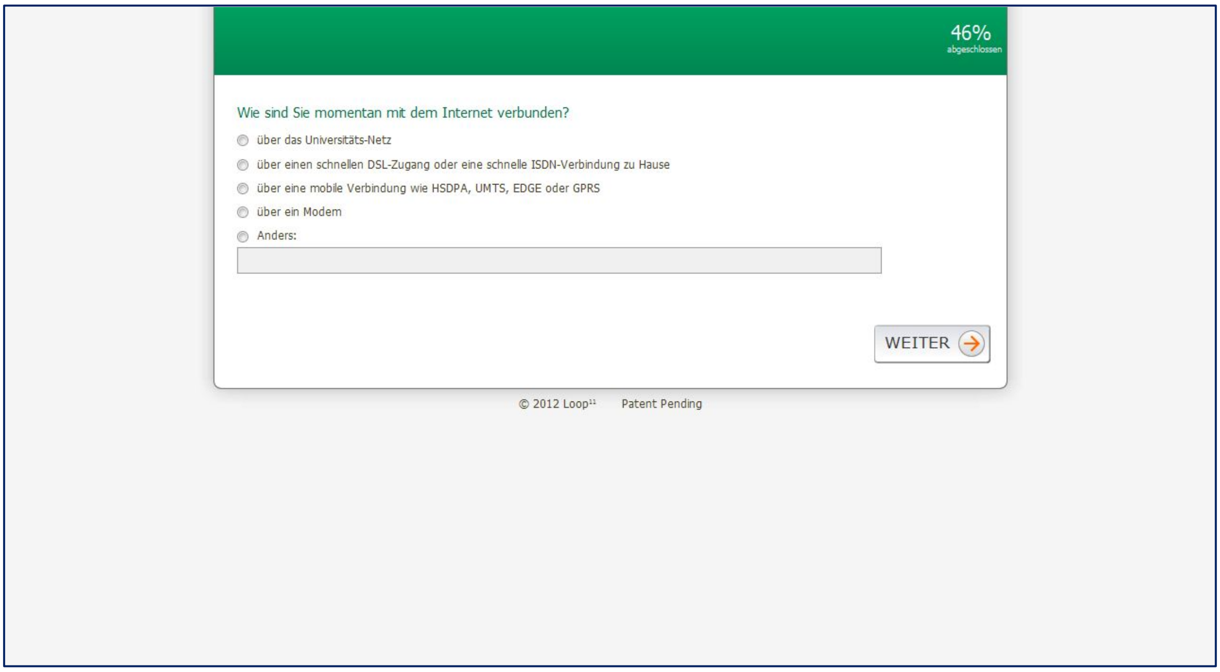
The Valley Project details life in two American communities, one Northern and one Southern, from the time of John Brown's Raid through the era of Reconstruction. In this digital archive you may explore thousands of original letters and diaries, newspapers and speeches, census and church records, left by men and women in Augusta County, Virginia, and Franklin County, Pennsylvania. Giving voice to hundreds of individual people, the Valley Project tells forgotten stories of life during the era of the Civil War.

Figure 56. Pilot test: Fifth task in Valley of the Shadow.



Sie haben diese Studie in einem Browser bearbeitet (z.B, Firefox oder Internet Explorer). Hatten Sie gleichzeitig andere Programme und Websites geöffnet (Emailprogramm, Facebook, Skype..)? Wenn Sie andere Programme geöffnet hatten: Welche? Bitte listen Sie diese einzeln auf:

Figure 57. Pilot test: Question about additional open programs.



Wie sind Sie momentan mit dem Internet verbunden?

○ über das Universitäts-Netz

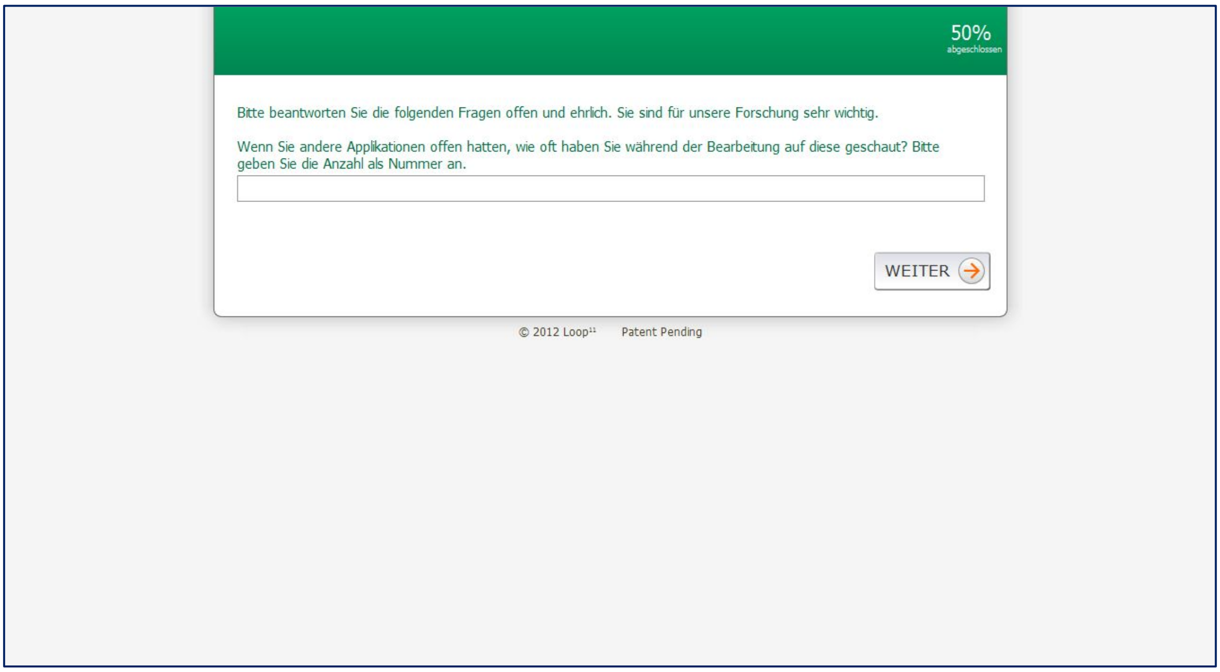
○ über einen schnellen DSL-Zugang oder eine schnelle ISDN-Verbindung zu Hause

○ über eine mobile Verbindung wie HSDPA, UMTS, EDGE oder GPRS

○ über ein Modem

○ Anders:

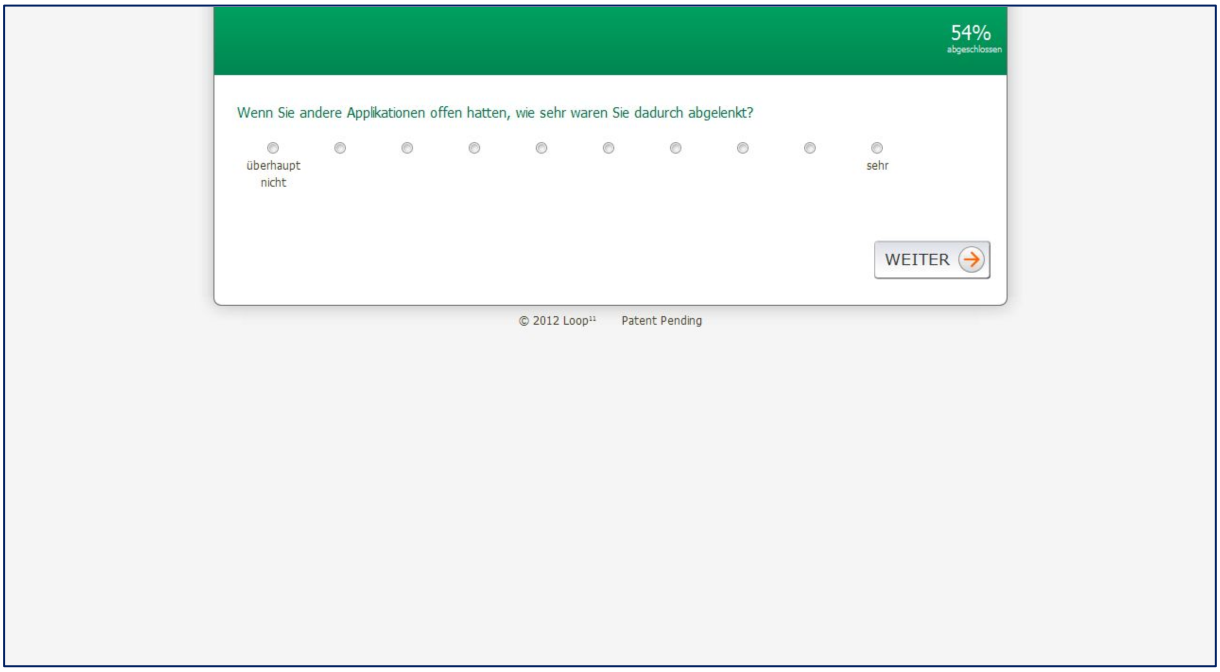
Figure 58. Pilot test: Question about participant's kind of internet connection.



Bitte beantworten Sie die folgenden Fragen offen und ehrlich. Sie sind für unsere Forschung sehr wichtig.

Wenn Sie andere Applikationen offen hatten, wie oft haben Sie während der Bearbeitung auf diese geschaut? Bitte geben Sie die Anzahl als Nummer an.

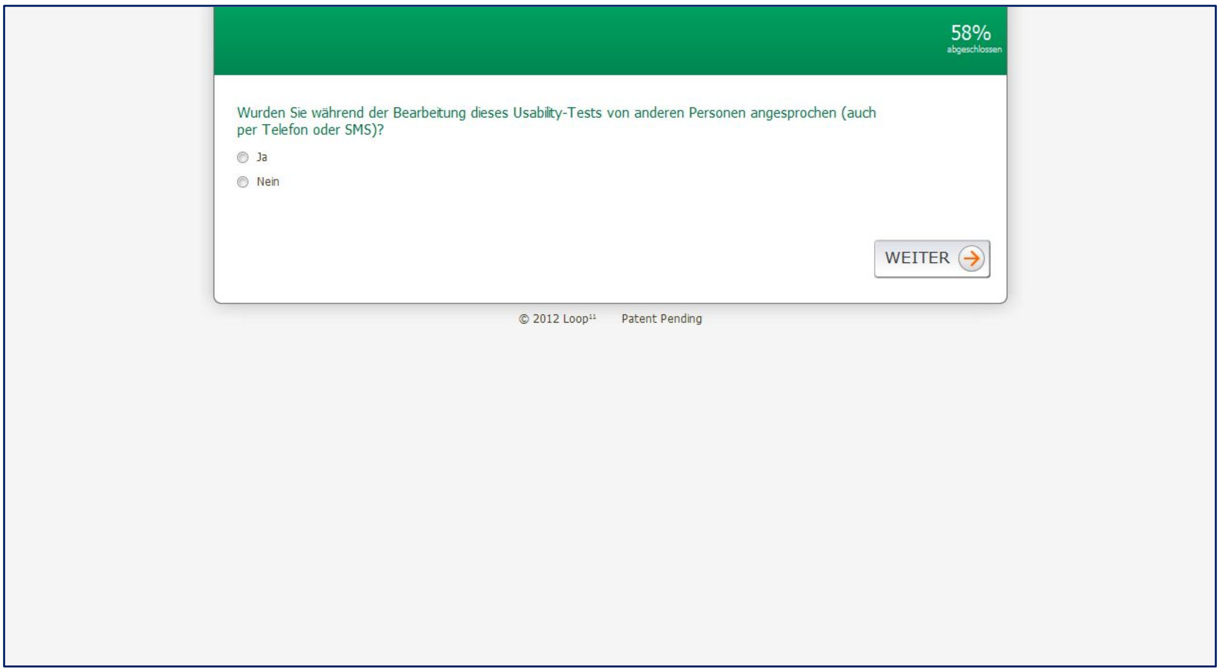
Figure 59. Pilot test: Information that honest answers about context information is important for this research. Question about the frequency of looking at the open programs.



Wenn Sie andere Applikationen offen hatten, wie sehr waren Sie dadurch abgelenkt?

○ überhaupt nicht ○ ○ ○ ○ ○ ○ ○ ○ ○ sehr

Figure 60. Pilot test: Question about the estimated level of distraction due to open programs.

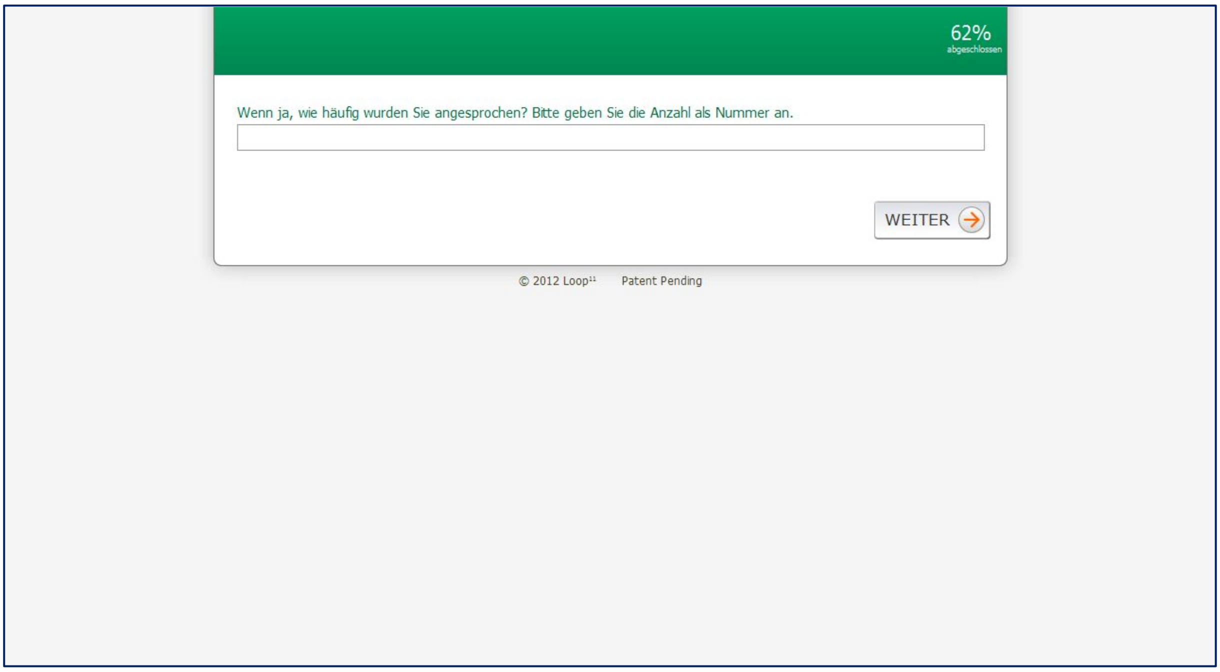


Wurden Sie während der Bearbeitung dieses Usability-Tests von anderen Personen angesprochen (auch per Telefon oder SMS)?

○ Ja

○ Nein

Figure 61. Pilot test: Question about potential contacts during the test.

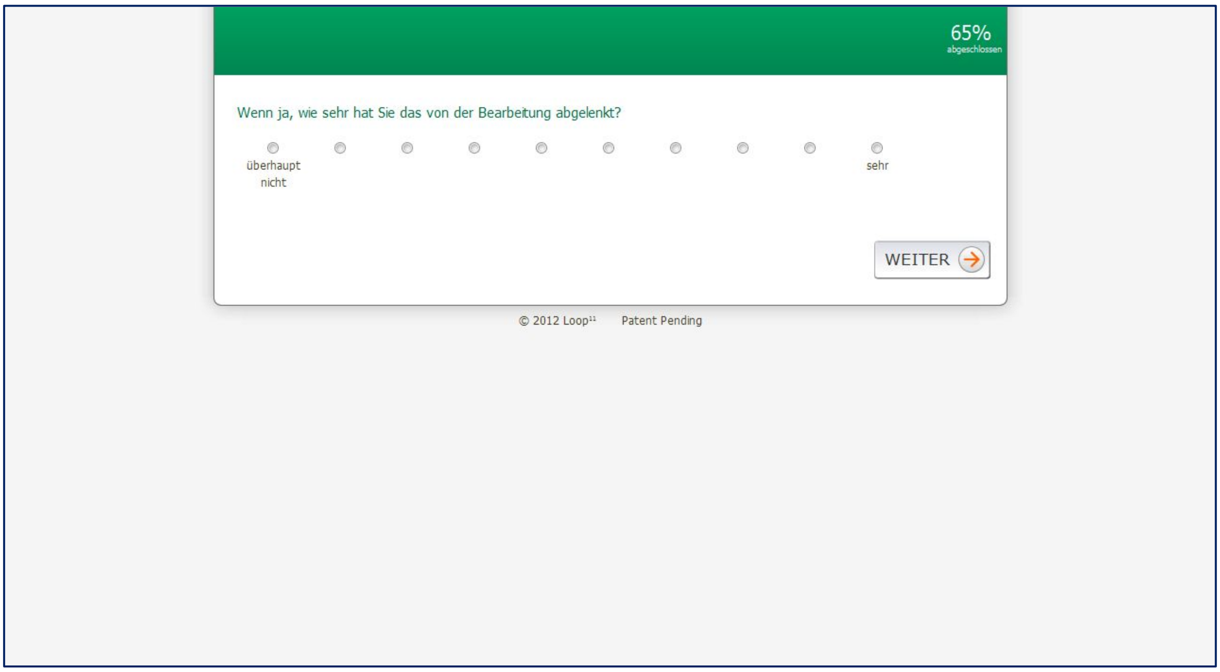


62%
abgeschlossen

Wenn ja, wie häufig wurden Sie angesprochen? Bitte geben Sie die Anzahl als Nummer an.

© 2012 Loop11 Patent Pending

Figure 62. Pilot test: Question about the frequency of being contacted.



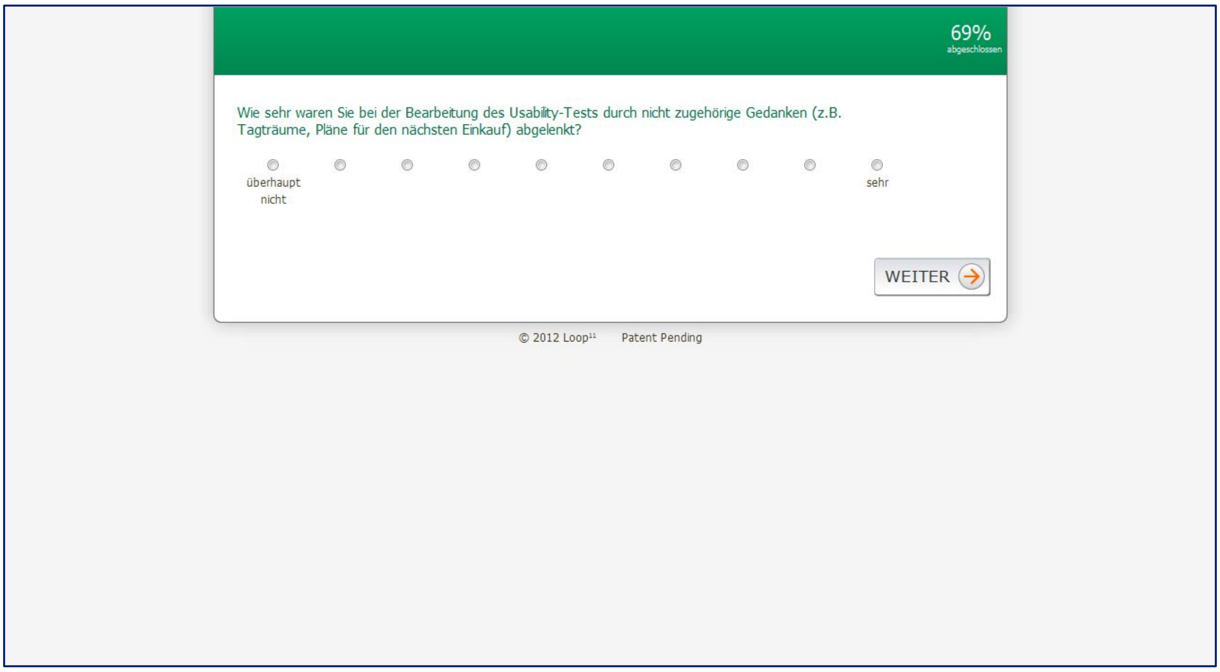
65%
abgeschlossen

Wenn ja, wie sehr hat Sie das von der Bearbeitung abgelenkt?

○ überhaupt nicht ○ ○ ○ ○ ○ ○ ○ ○ ○ sehr

© 2012 Loop11 Patent Pending

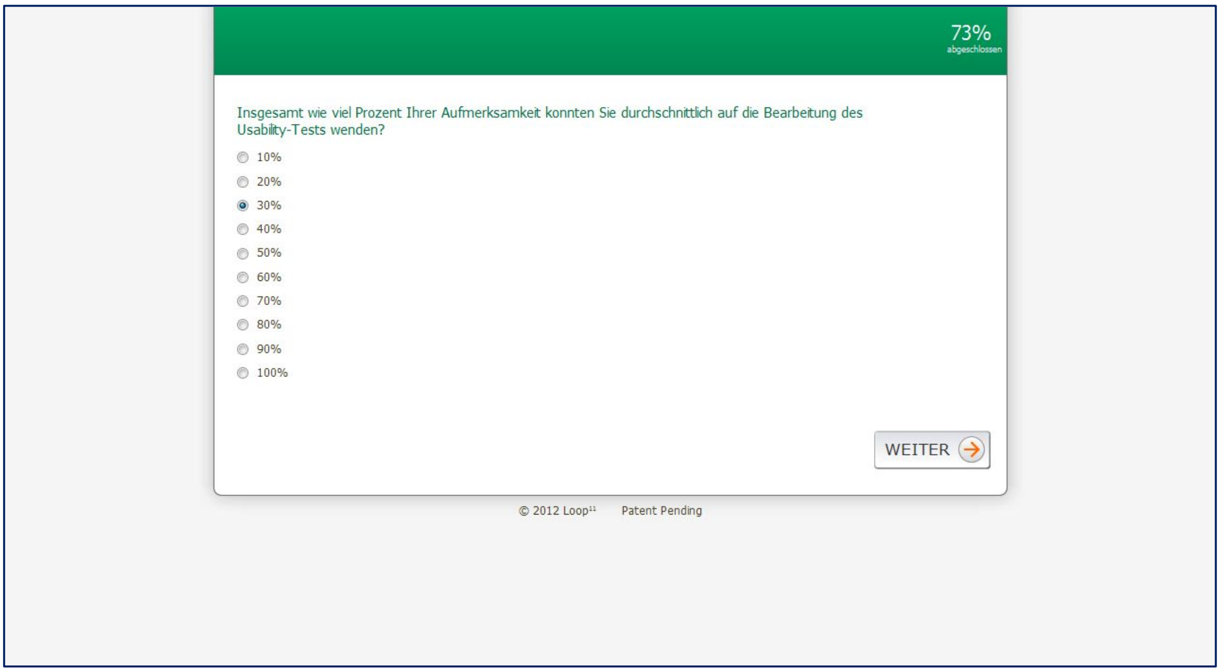
Figure 63. Pilot test: Question about the estimated level of distraction due to contacts.



Wie sehr waren Sie bei der Bearbeitung des Usability-Tests durch nicht zugehörige Gedanken (z.B. Tagträume, Pläne für den nächsten Einkauf) abgelenkt?

○	○	○	○	○	○	○	○	○	○
überhaupt nicht									sehr

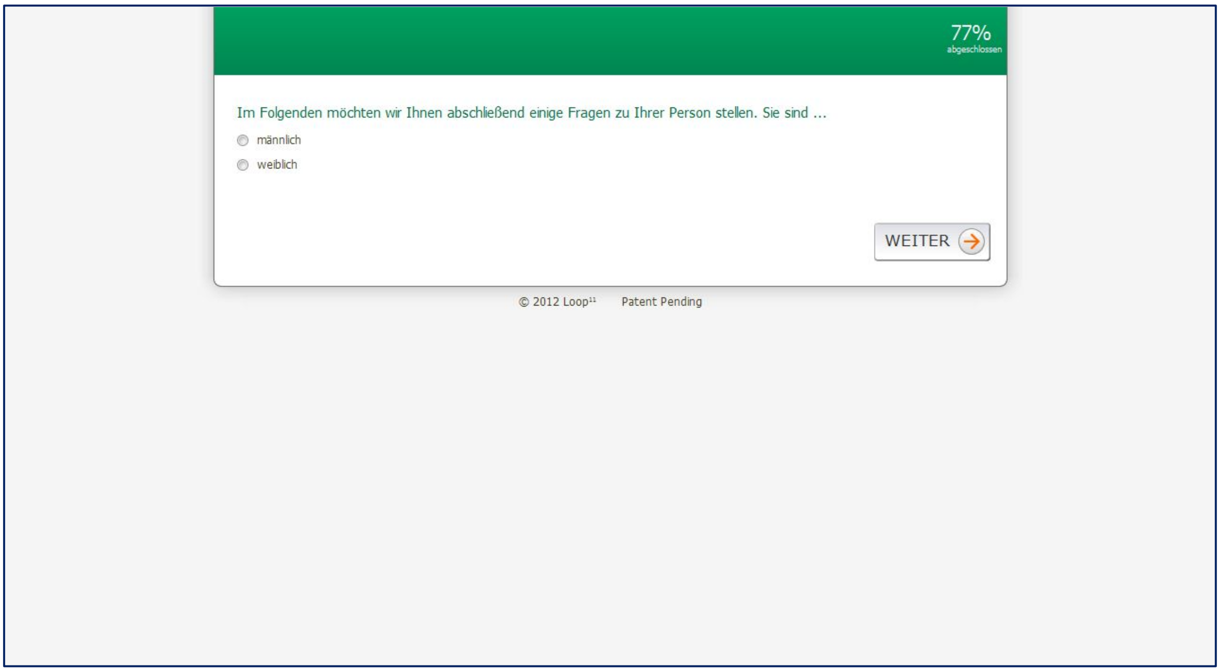
Figure 64. Pilot test: Question about the estimated level of distraction due to daydreams.



Insgesamt wie viel Prozent Ihrer Aufmerksamkeit konnten Sie durchschnittlich auf die Bearbeitung des Usability-Tests wenden?

- ○ 10%
- ○ 20%
- ● 30%
- ○ 40%
- ○ 50%
- ○ 60%
- ○ 70%
- ○ 80%
- ○ 90%
- ○ 100%

Figure 65. Pilot test: Question about participant's level of attention on the test in percentages.



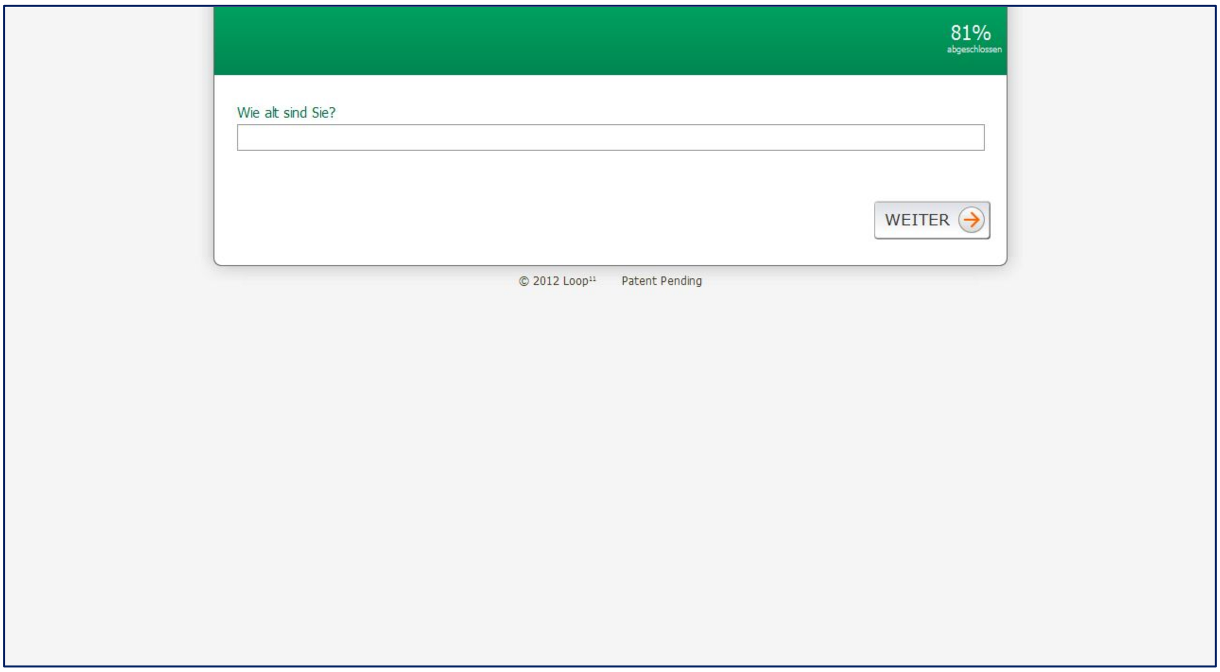
77%
abgeschlossen

Im Folgenden möchten wir Ihnen abschließend einige Fragen zu Ihrer Person stellen. Sie sind ...

○ männlich

○ weiblich

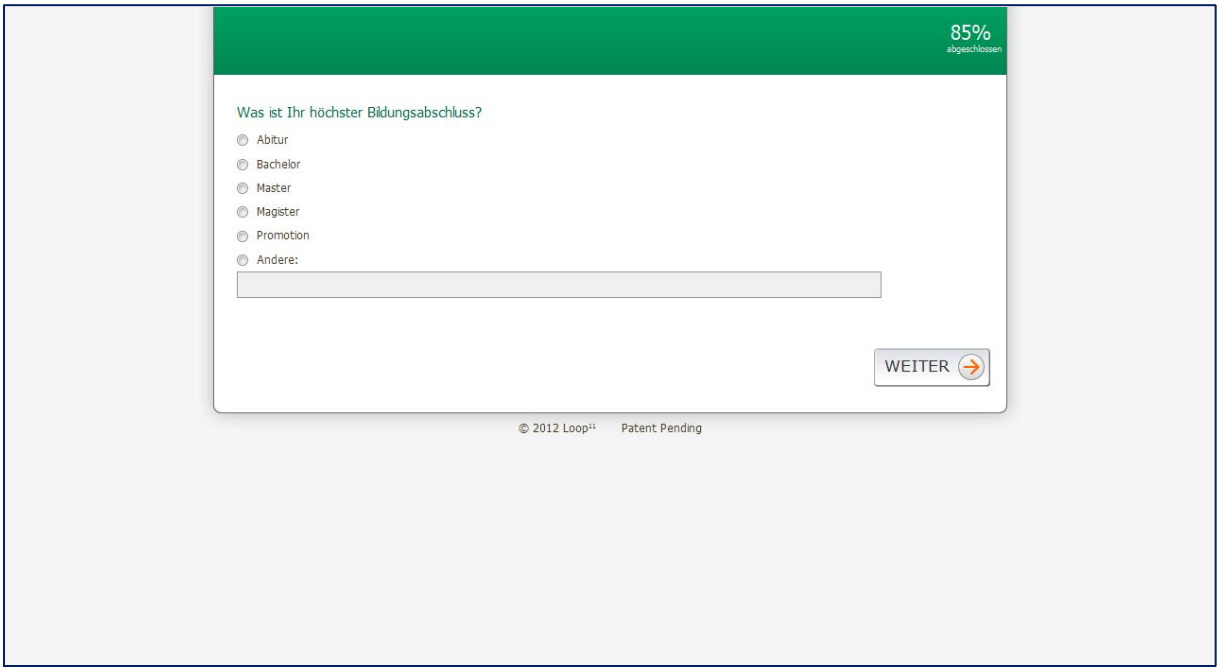
Figure 66. Pilot test: Question about the participant’s gender.



81%
abgeschlossen

Wie alt sind Sie?

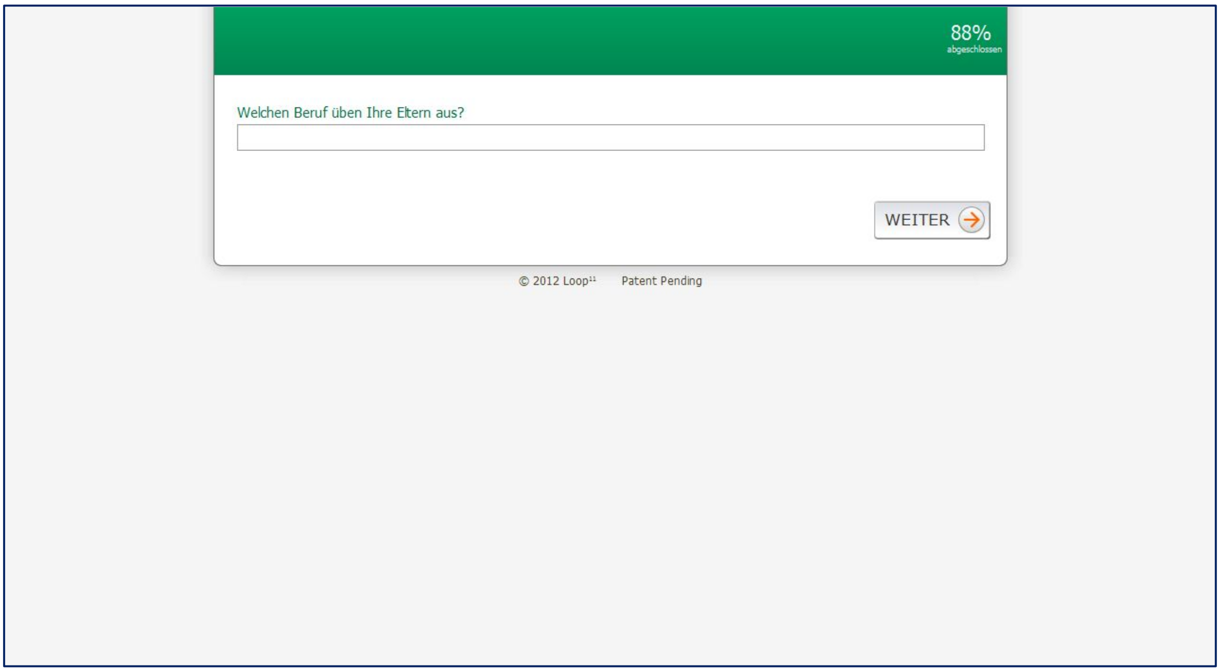
Figure 67. Pilot test: Question about the participant’s age.



Was ist Ihr höchster Bildungsabschluss?

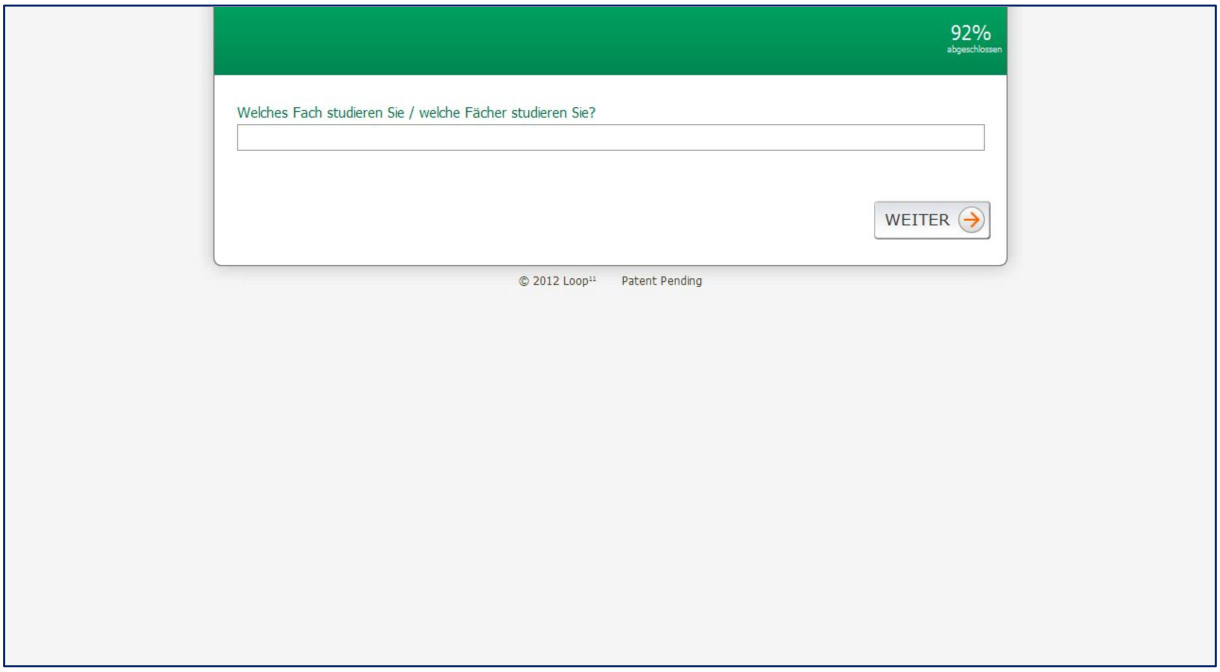
○ Abitur
○ Bachelor
○ Master
○ Magister
○ Promotion
○ Andere:

Figure 68. Pilot test: Question about the participant's highest university degree.



Welchen Beruf üben Ihre Eltern aus?

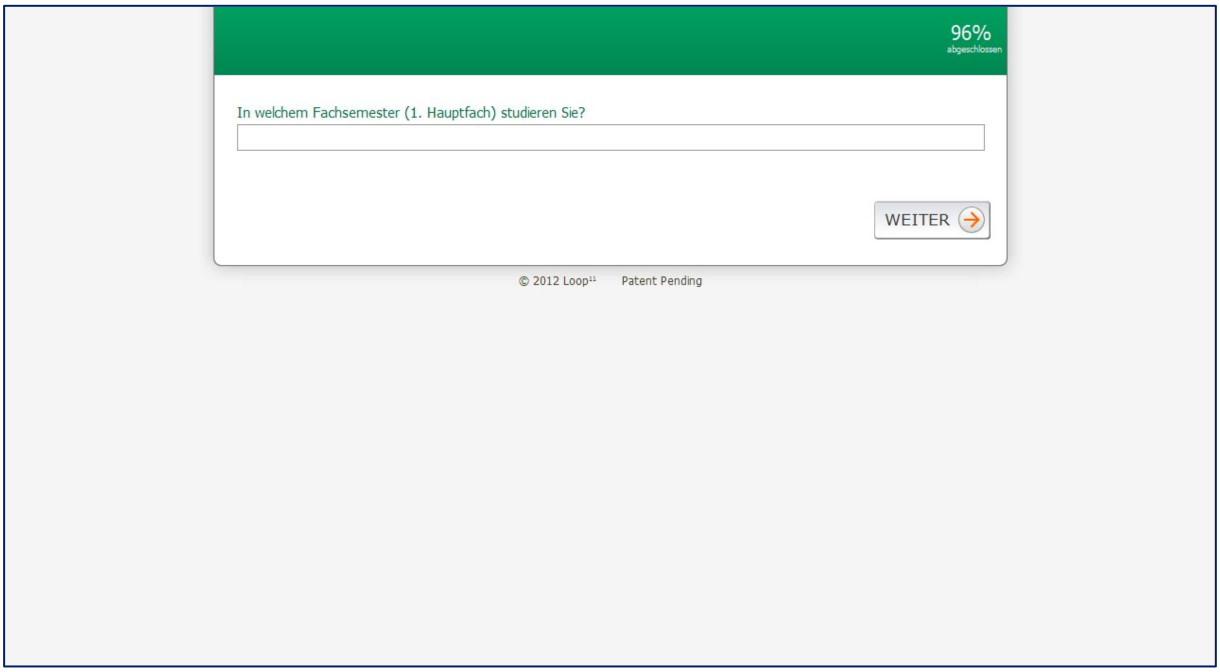
Figure 69. Pilot test: Question about the job the participant's parents hold.



92%
abgeschlossen

Welches Fach studieren Sie / welche Fächer studieren Sie?

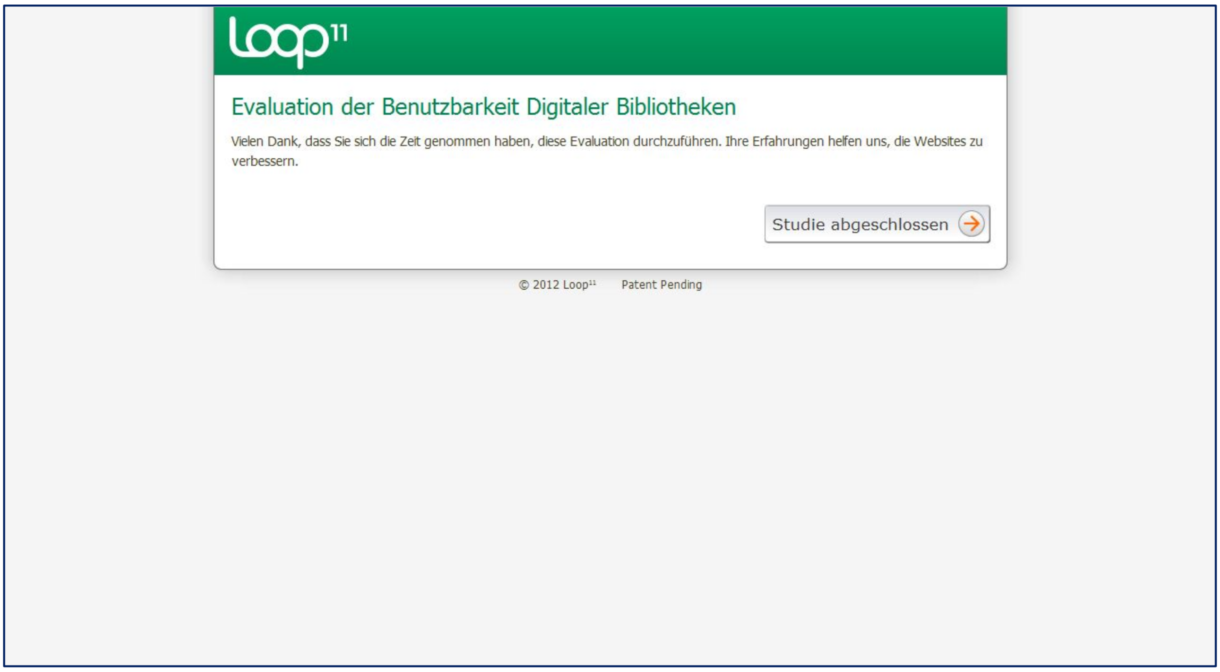
Figure 70. Pilot test: Question about the participant's academic subject.



96%
abgeschlossen

In welchem Fachsemester (1. Hauptfach) studieren Sie?

Figure 71. Pilot test: Question about the participant's current number of semesters.



Evaluation der Benutzbarkeit Digitaler Bibliotheken

Vielen Dank, dass Sie sich die Zeit genommen haben, diese Evaluation durchzuführen. Ihre Erfahrungen helfen uns, die Websites zu verbessern.

Figure 72. Pilot test: Good-bye page thanking the participant.

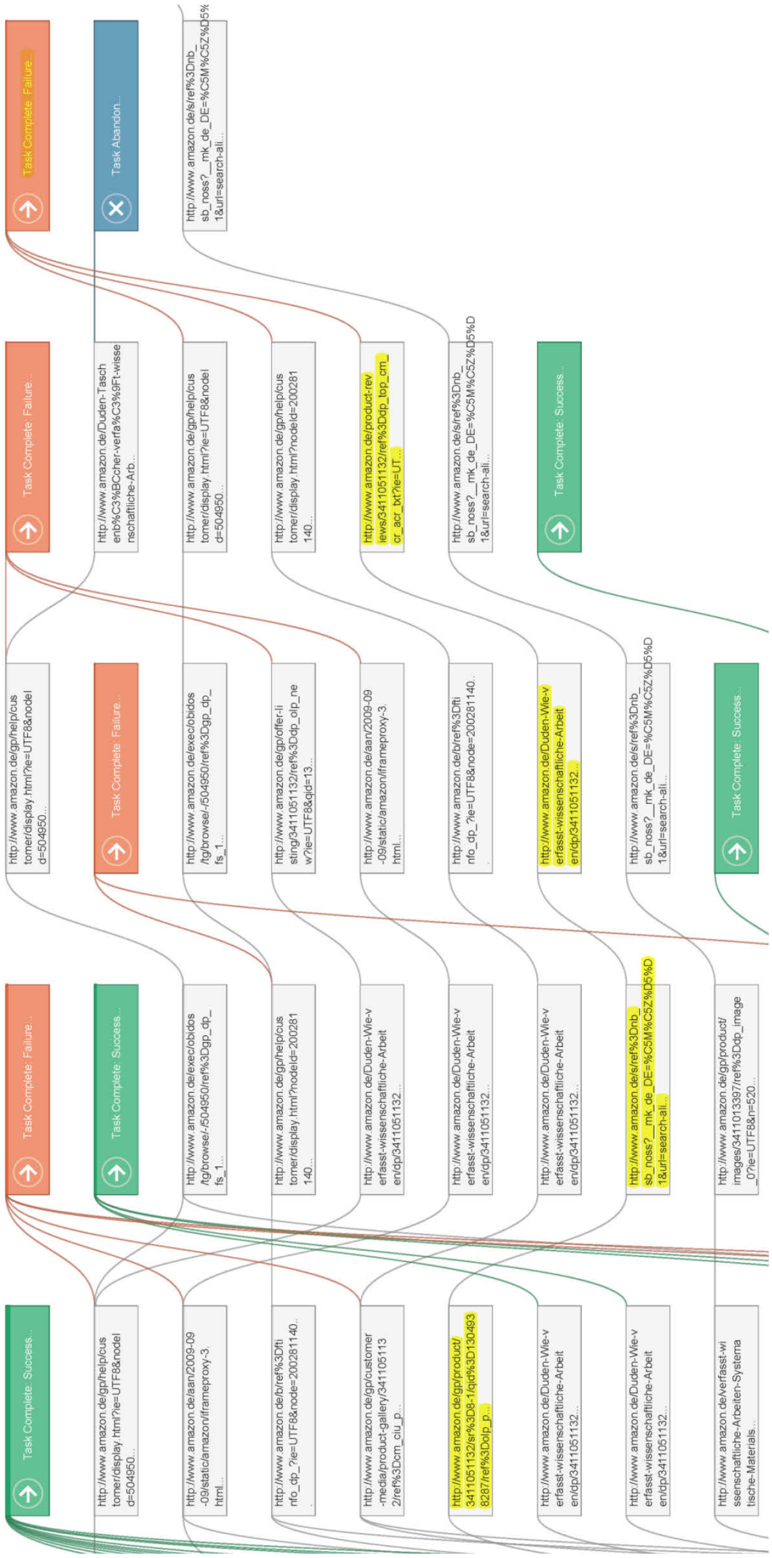
Appendix 5. Academic specializations

The test asked participants to name their first, second or third academic specializations. The specializations below are original labels by the participants. Sometimes subjects in Germany have a different orientation than in the United States. Art history, for example, is actually a history of art. Musicology or dramatics at a university is the study of the development of music or theatre. Studying music or theatre in the sense of making music or playing theatre is mostly done at special schools.

1 = Text oriented subjects	Ägyptologie	Medienwissenschaft
	Anglistik	Musikwissenschaft
	Deutsch	Neuere deutsche Geschichte
	Europäische Ethnologie	Philosophie
	Europäische Studien	Politikwissenschaft
	Europastudien	Portugiesisch
	Filmwissenschaft	Publizistik
	Gender Studies	Regionalstudien
	Germanistik	Regionalstudien Asien und Afrika
	Geschichtswissenschaft	Schauspiel
	Jura	Sozialkunde
	Kommunikationswissenschaft	Sprachwissenschaft
	Kulturmanagement	Theaterwissenschaft
	Kulturtourismus	Theologie
	Kulturwissenschaften	Urban Sociology
	Kunstgeschichte	Wirtschaftspolitik
	Literaturwissenschaft	
2 = Mathematically oriented subjects	Betriebswirtschaftslehre	Medizinpädagogik
	Biologie	Metallurgie
	Chemie	Physik
	Geographie	Rehabilitationspädagogik
	Geophysik	Rehabilitationswissenschaft
	Marketing	Sportwissenschaft
	Mathematik	Umweltwissenschaften
	Medieninformatik	Volkswirtschaftslehre
	Mediensysteme (angewandte Informatik)	Wirtschaftsinformatik
	Medizin	Wirtschaftswissenschaft
3 = Subjects with test design experience	Bibliotheks- und Informationswissenschaft	Sozialarbeit
	Erziehungswissenschaft	Sozialwissenschaften
	Grundschulpädagogik	Soziologie
	Psychologie	

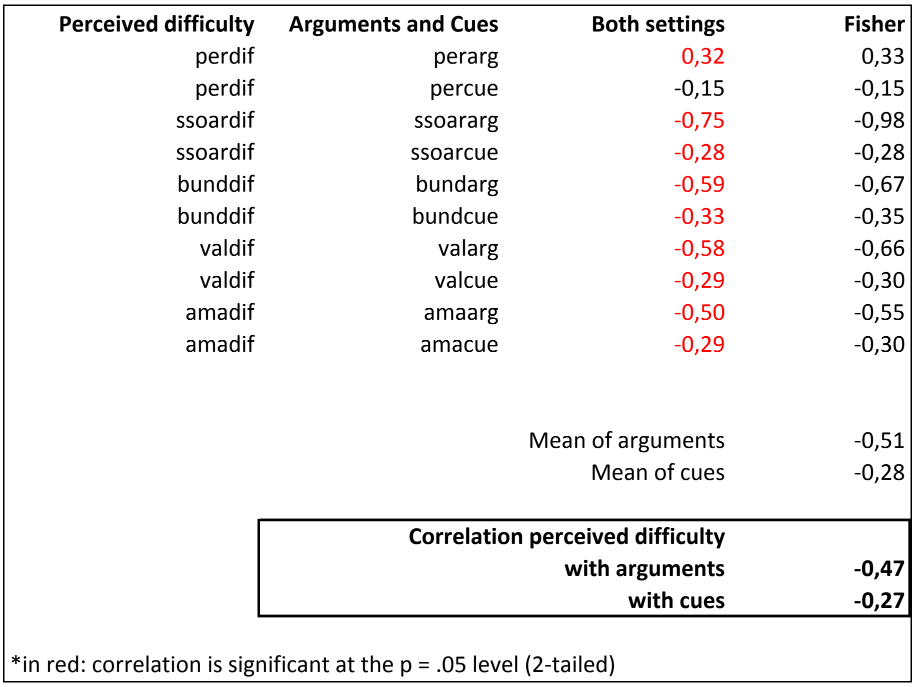
Appendix 6. Participants' clickstream from the task using Amazon

Appendix 6 exemplifies the behavior of one participant doing the task in *Amazon* (highlighted in yellow) at the stage of three previous page views.



Task Complete: Failure...
Task Abandon...
http://www.amazon.de/s/ref%3Dnb_sb_noss?__mk_de_DE=%C5M%C5Z%D5%1&url=search-ali...
Task Complete: Failure...
http://www.amazon.de/Duden-Taschenb%C3%BCcher-verfa%C3%9Ft-wissenschaftliche-Arb...
http://www.amazon.de/gp/help/customer/display.html?ie=UTF8&nodeId=504950...
http://www.amazon.de/gp/help/customer/display.html?nodeId=200281140...
http://www.amazon.de/product-reviews/3411051132/ref%3Ddp_top_cm_cr_acr_txt?ie=UT...
http://www.amazon.de/s/ref%3Dnb_sb_noss?__mk_de_DE=%C5M%C5Z%D5%D1&url=search-ali...
Task Complete: Success...
http://www.amazon.de/gp/help/customer/display.html?ie=UTF8&nodeId=504950...
Task Complete: Failure...
http://www.amazon.de/exec/obidos/tg/browse/-/504950/ref%3Dgp_dp_fs_1...
http://www.amazon.de/gp/offer-listing/3411051132/ref%3Ddp_olp_new?ie=UTF8&qid=13...
http://www.amazon.de/aan/2009-09/static/amazon/iframeproxy-3.html...
http://www.amazon.de/b/ref%3Dfooter_info_dp_?ie=UTF8&node=200281140...
http://www.amazon.de/Duden-Wie-verfasst-wissenschaftliche-Arbeiten/dp/3411051132...
http://www.amazon.de/s/ref%3Dnb_sb_noss?__mk_de_DE=%C5M%C5Z%D5%D1&url=search-ali...
Task Complete: Success...
Task Complete: Failure...
Task Complete: Success...
http://www.amazon.de/exec/obidos/tg/browse/-/504950/ref%3Dgp_dp_fs_1...
http://www.amazon.de/gp/help/customer/display.html?nodeId=200281140...
http://www.amazon.de/Duden-Wie-verfasst-wissenschaftliche-Arbeiten/dp/3411051132...
http://www.amazon.de/Duden-Wie-verfasst-wissenschaftliche-Arbeiten/dp/3411051132...
http://www.amazon.de/Duden-Wie-verfasst-wissenschaftliche-Arbeiten/dp/3411051132...
http://www.amazon.de/s/ref%3Dnb_sb_noss?__mk_de_DE=%C5M%C5Z%D5%D1&url=search-ali...
http://www.amazon.de/gp/product/images/3411013397/ref%3Ddp_image_0?ie=UTF8&n=520...
Task Complete: Success...
http://www.amazon.de/gp/help/customer/display.html?ie=UTF8&nodeId=504950...
http://www.amazon.de/aan/2009-09/static/amazon/iframeproxy-3.html...
http://www.amazon.de/b/ref%3Dfooter_info_dp_?ie=UTF8&node=200281140...
http://www.amazon.de/gp/customer-media/product-gallery/3411051132/ref%3Dcm_ciu_p...
http://www.amazon.de/gp/product/3411051132/sr%3D8-1/qid%3D1304938287/ref%3Dolp_p...
http://www.amazon.de/Duden-Wie-verfasst-wissenschaftliche-Arbeiten/dp/3411051132...
http://www.amazon.de/Duden-Wie-verfasst-wissenschaftliche-Arbeiten/dp/3411051132...
http://www.amazon.de/verfasst-wissenschaftliche-Arbeiten-Systematische-Materials...

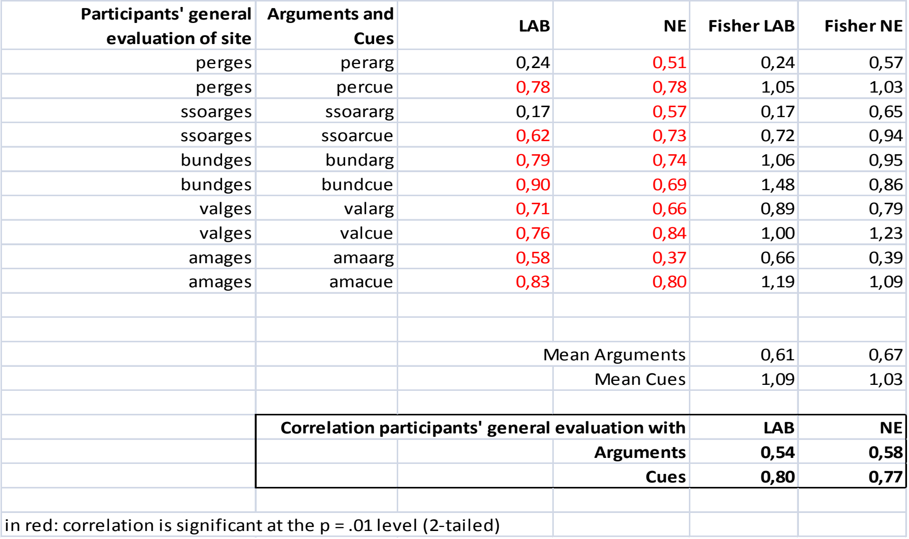
Appendix 7. Correlations for an assessment of the decision-making processes



Perceived difficulty	Arguments and Cues	Both settings	Fisher
perdif	perarg	0,32	0,33
perdif	percue	-0,15	-0,15
ssoardif	ssoararg	-0,75	-0,98
ssoardif	ssoarcue	-0,28	-0,28
bunddif	bundarg	-0,59	-0,67
bunddif	bundcue	-0,33	-0,35
valdif	valarg	-0,58	-0,66
valdif	valcue	-0,29	-0,30
amadif	amaarg	-0,50	-0,55
amadif	amacue	-0,29	-0,30
		Mean of arguments	-0,51
		Mean of cues	-0,28
		Correlation perceived difficulty	
		with arguments	**-0,47**
		with cues	**-0,27**

*in red: correlation is significant at the p = .05 level (2-tailed)

Table 14. Correlations between perceived difficulty and arguments, and between perceived difficulty and cues in both settings.

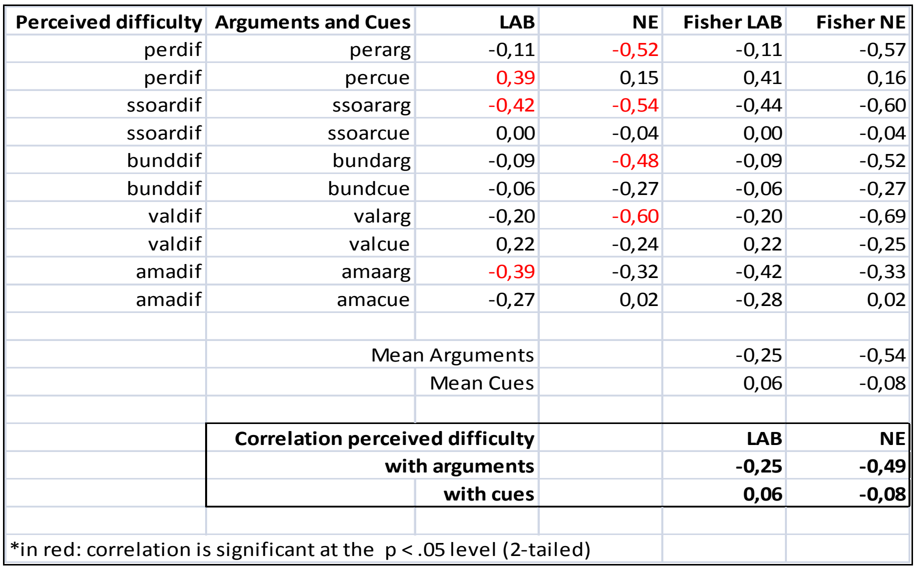


Participants' general evaluation of site	Arguments and Cues	LAB	NE	Fisher LAB	Fisher NE
perges	perarg	0,24	0,51	0,24	0,57
perges	percue	0,78	0,78	1,05	1,03
ssoarges	ssoararg	0,17	0,57	0,17	0,65
ssoarges	ssoarcue	0,62	0,73	0,72	0,94
bundges	bundarg	0,79	0,74	1,06	0,95
bundges	bundcue	0,90	0,69	1,48	0,86
valges	valarg	0,71	0,66	0,89	0,79
valges	valcue	0,76	0,84	1,00	1,23
amages	amaarg	0,58	0,37	0,66	0,39
amages	amacue	0,83	0,80	1,19	1,09
			Mean Arguments	0,61	0,67
			Mean Cues	1,09	1,03
			Correlation participants' general evaluation with	**LAB**	**NE**
			Arguments	**0,54**	**0,58**
			Cues	**0,80**	**0,77**

in red: correlation is significant at the p = .01 level (2-tailed)

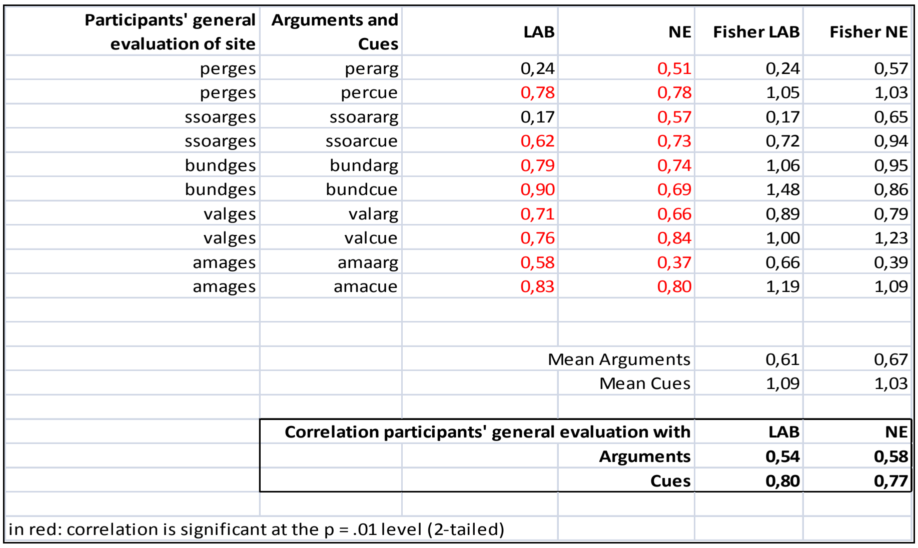
<

Table 15. Correlations between participants' general evaluation of a site and arguments, and participants' general evaluation of a site and cues in both settings.



Perceived difficulty	Arguments and Cues	LAB	NE	Fisher LAB	Fisher NE
perdif	perarg	-0,11	-0,52	-0,11	-0,57
perdif	percue	0,39	0,15	0,41	0,16
ssoardif	ssoararg	-0,42	-0,54	-0,44	-0,60
ssoardif	ssoarcue	0,00	-0,04	0,00	-0,04
bunddif	bundarg	-0,09	-0,48	-0,09	-0,52
bunddif	bundcue	-0,06	-0,27	-0,06	-0,27
valdif	valarg	-0,20	-0,60	-0,20	-0,69
valdif	valcue	0,22	-0,24	0,22	-0,25
amadif	amaarg	-0,39	-0,32	-0,42	-0,33
amadif	amacue	-0,27	0,02	-0,28	0,02
		Mean Arguments		-0,25	-0,54
		Mean Cues		0,06	-0,08
		Correlation perceived difficulty		**LAB**	**NE**
		with arguments		**-0,25**	**-0,49**
		with cues		**0,06**	**-0,08**
*in red: correlation is significant at the p < .05 level (2-tailed)					

Table 16. Correlations between perceived difficulty and arguments, and between perceived difficulty and cues in the laboratory and natural environments.



Participants' general evaluation of site	Arguments and Cues	LAB	NE	Fisher LAB	Fisher NE
perges	perarg	0,24	0,51	0,24	0,57
perges	percue	0,78	0,78	1,05	1,03
ssoarges	ssoararg	0,17	0,57	0,17	0,65
ssoarges	ssoarcue	0,62	0,73	0,72	0,94
bundges	bundarg	0,79	0,74	1,06	0,95
bundges	bundcue	0,90	0,69	1,48	0,86
valges	valarg	0,71	0,66	0,89	0,79
valges	valcue	0,76	0,84	1,00	1,23
amages	amaarg	0,58	0,37	0,66	0,39
amages	amacue	0,83	0,80	1,19	1,09
			Mean Arguments	0,61	0,67
			Mean Cues	1,09	1,03
			Correlation participants' general evaluation with	**LAB**	**NE**
			Arguments	**0,54**	**0,58**
			Cues	**0,80**	**0,77**
in red: correlation is significant at the p = .01 level (2-tailed)					

Table 17. Correlation between participants' general evaluation with arguments and participants' general evaluation with cues in the laboratory and natural environments.

Participants' general evaluation of site	Arguments and Cues	With programs open	No programs open	Fisher with programs	Fisher no programs
perges	perarg	0,16	0,44	0,16	0,47
perges	percue	0,79	0,76	1,08	1,00
ssoarges	ssoararg	0,45	0,42	0,48	0,44
ssoarges	ssoarcue	0,63	0,71	0,74	0,89
bundges	bundarg	0,72	0,72	0,90	0,90
bundges	bundcue	0,77	0,86	1,02	1,29
valges	valarg	0,66	0,70	0,80	0,87
valges	valcue	0,92	0,72	1,59	0,90
amages	amaarg	0,42	0,54	0,45	0,60
amages	amacue	0,82	0,82	1,16	1,14
			Mean Arguments	0,56	0,66
			Mean Cues	1,12	1,05
	Correlation participants' general evaluation with			Program open	No programs
			Arguments	0,51	0,58
			Cues	0,81	0,78
*in red correlation is significant at the 0.01 level (2-tailed)					

Table 18. Correlation between participants' general evaluation with arguments and participants' general evaluation with cues, in groups with participants who had programs open and groups without open programs.

Participants' general evaluation of site	Arguments and Cues	Contacted	No contact	Fisher contacted	Fisher no contact
perges	perarg	0,37	0,32	0,39	0,33
perges	percue	0,85	0,77	1,26	1,02
ssoarges	ssoararg	0,39	0,43	0,41	0,45
ssoarges	ssoarcue	0,86	0,63	1,29	0,74
bundges	bundarg	0,74	0,79	0,95	1,06
bundges	bundcue	0,83	0,85	1,20	1,27
valges	valarg	0,48	0,71	0,52	0,89
valges	valcue	0,83	0,79	1,19	1,06
amages	amaarg	0,59	0,51	0,67	0,57
amages	amacue	0,66	0,84	0,78	1,22
			Mean Arguments	0,59	0,66
			Mean Cues	1,15	1,06
	Correlation participants' general evaluation with			Contacted	No contact
			Arguments	0,53	0,58
			Cues	0,82	0,79
in red correlation is significant at the 0.01 level (2-tailed)					

Table 19. Correlation between participants' general evaluation with arguments and participants' general evaluation with cues, in groups with participants who were contacted and groups with participants who had not been contacted during the test.