Assessment of Usability Benchmarks

Combining Standardized Scales with Specific Questions

http://dx.doi.org/10.3991/ijet.v6i4.1832

Stephanie B. Linek & Klaus Tochtermann
ZBW – Leibniz Information Center of Economics, Kiel, Germany

***Abstract*—The usability of Web sites and online services is of rising importance. When creating a completely new Web site, qualitative data are adequate for identifying the most usability problems. However, changes of an existing Web site should be evaluated by a quantitative benchmarking process. The proposed paper describes the creation of a questionnaire that allows a quantitative usability benchmarking, i.e. a direct comparison of the different versions of a Web site and an orientation on general standards of usability. The questionnaire is also open for qualitative data. The methodology will be explained by the digital library services of the ZBW.**

***Index Terms*—usability evaluation, assessment, methodology, benchmarking questionnaire**

I. USABILITY EVALUATION OF WEB SITES AND ONLINE SERVICES

The usability of Web sites and online services is of rising importance. Thereby, it is often claimed, that qualitative data are sufficient for usability evaluation. But this is only half of the story. Indeed, qualitative data are adequate for identifying the most usability problems in a rough way. However, in some cases it is necessary to go a step further and make the usability level and improvements quantifiable.

For the usability evaluation of Web sites one can differentiate between two main scenarios: First, the Web sites and the services do not yet exist and have to be constructed. Second, an existing Web site with its well-established services has to be improved or enriched with new features. In the first case it makes sense to work with qualitative data and small groups of people to identify main problems in advance or receive information about the requirements of end-users. At very early stages this can done for example by the help of paper-based mock-ups. At later stages, a clickable incomplete version of the planned Web site (i.e., without the full functionalities) can simulate the prospective human-computer interaction. In the second case, the situation looks a bit different. Instead of developing a complete new Web site, the existing Web site has to be enhanced and improved. This in turn means that there is a starting point (baseline) as well as an existing group of end-users. Smaller changes could be well evaluated by qualitative data. However, when there are substantial changes, it is more appropriate to make a kind of benchmarking measurement with a larger group of users and well defined quantitative measurement instruments. Even though the planned changes are thoroughly created and implemented, this doesn't guarantee a change for the better. Thus, it has to be evaluated if the changes are actual an improvement for the end-users.

The prerequisite for a quantitative benchmark is to have quantitative measurements of usability. On first sight, that seems trivial. However, in practice several aspects might be problematic or complicated. Particularly, the following points are decisional: It has to be defined what is meant by "usability", an appropriate measurement instrument has to be chosen, and it has to be clarified, which level of usability is good enough. These aspects are closely interconnected and depend from each other.

In literature there are several slightly different definitions of usability. Besides accessibility, most definitions refer to four aspects: Effectiveness, efficiency, appropriateness for subjective aims of the user (usefulness), and joy of use (see for example [1] or the International Standard Organization [2]). For the usability of Web sites and online services, the design of the human-computer interface is of special importance. Thereby, often mentioned additional characteristics of good usability are error recovery, easy to learn, easy to remember, error tolerant and likeability (examples can be found in [3], [4], [5]; see also discussion by [6]). Besides the general definition of usability one has to decide if there are specific aspects or indicators of good usability in the concrete case, for example, the content quality of a literature data base (as indicator for efficiency) or the flow-experience for a serious game (as indicator for joy of use).

After defining what is meant by usability and which aspects are the most important ones, the question arises, how much usability is good enough – or in other words: Where is the benchmark. A benchmark can be defined as a point of reference or a standard, that allows a comparison or a judgment how good or bad other things are. This definition illustrates two important points: First, it has to be decided which kind of measurement is appropriate, e.g. a global (subjective) rating or the objective (behavioral) error rate. Second, the quantitative level that constitutes the comparison standard has to be defined. Both aspects of an appropriate benchmark depend on the goal of usability evaluation and the concrete application case. If you are lucky, you can regard to an existing usability benchmark. For example, you know that users tolerate a maximum of two bugs without frustration. Or you know the usability scores of a comparable leading software and the accordingly scale is available (and usable) for your purpose. But that's not the usual case. Normally, competitors don't give away their usability results. Additionally, software products and web sites are quite different and normally one is not only interested in a global rating but also in a specific benchmark for the unique characteristics of a Web site or software. For example, one can compare the usability of gmx-email and Google search on a global level, but not for single features, because the both services

address different purposes. For gmx-email you might be interested in the usability of text editing and easiness of downloading the attachment of an email. For Google search it is more relevant to have usability rating regarding the handling and the quality of the search results. On this fine-grained level it makes no sense to compare the usability of text editing for GMX-email with the usability of handling the search result list of Google.

If there are no existing benchmarks available, you can either set a desired standard, e.g. getting always the highest possible scores of a usability rating – which is quite utopistic – or you can create your own realistic benchmark by measuring the status quo. The measured status quo can be conceptualized as the baseline-standard, i.e. the starting point. After innovations, changes or a re-design, the measurement can be repeated and the result can be compared with the original baseline-data (standard). That means, this enables a direct comparison between the usability level before and after the changes. Depending on the concrete measurement instrument, such kind of benchmarking enables also a comparison with other Web sites or services. It's important to note, that even without any changes it could be necessary to make a repeated usability benchmarking since the expectations of the end-users could change, for example they expect other functions due to the progress of technology.

In general, for the measurement of usability different methods are thinkable: Usability-testing, questionnaires or heuristic evaluation by experts (overview is given by [1]). However, not all methods are equal appropriate for usability benchmarking. To set a quantitative standard for usability (or in other words a usability benchmark), we need *quantitative* measurements from an *appropriate sample* of user. Thus, methods like heuristic evaluation that deliver mainly qualitative (non-empirical) data are less apt. Usability testing in principle would work very well but is rather expensive when applied on a large sample of users. If one wants a representative benchmark assessed with a big sample of users, it is usually more appropriate to use a questionnaire that can be easily distributed among many users (normally without moderators). While usability tests deliver mainly behavioral data like error rate and time of task completion, questionnaires provide subjective ratings regarding ease of use or the overall impression. Both, behavioral data and subjective ratings have their advantages and drawbacks and it is beneficial to have a combination of both. An example of the combination of behavioral logfile-data and subjective data of questionnaires is given in [7]. However, such a combination is often cost-intensive. Thus, it has to be carefully decided, which kind of indicator is appropriate for the concrete purpose. The core question is: what are the most appropriate indicators for the benchmark - behavioral data or subjective ratings measured by a questionnaire?

Imagine for example special medical software for the application of the right doses of a medicine. In this case it is essential, that the software is safe and quick in handling. However, the joy of use is less important. Thus, the usability benchmark should de defined (at least partly) by the error rate and completion time for the functionalities that aim at life saving.

Contrariwise, the Web site of a game community, aims at a high level of joy of use and an attractive design that motivates the users for surfing around the Web site. Thus, the overall impression is more important than the error rate or completion time for single functionalities. Additionally, pure behavioral indicators are often ambiguous. For example, the duration time on the Web site comprises no distinct information: is it due to confusion or due to fascination? Thus, in the case of the Web site of a game community it might be more appropriate to use subjective ratings like the overall impression or the individual judgments of users regarding design, handling etc. Such indicators can be well measured by questionnaires.

After the selection of appropriate indicators (behavioral measures like error rate or subjective data like overall impression), one has to set a quantitative benchmark. What is good enough? And how good we are at the moment? In some cases, the benchmark is directly dependent on the application case. In the example of the medical software, the error rate has to be zero if an error would cause the death of the patient. Also the completion time has to be defined in the light of medical demands, e.g. the time a human being can stand without oxygen. However, this is a very extreme example. For a normal Web site or online service the usability benchmark is less obvious. As explained above, if no appropriate existing benchmark is available, a baseline measurement is often the most appropriate practical standard one can set.

In this paper we propose a benchmarking questionnaire for the measurement of subjective indicators that can be used to quantify the progress by means of repeated cycles of data assessment. The benchmarking questionnaire comprises standardized scales as well as specific questions. This method has several advantages:

- Standardized scales enable the comparison with general standards of usability, with competitors and with prior versions of a Web site on a general level.
- Specific questions can address the specific features of the Web site and the requirements of the user.
- Combined interpretation of standardized scales and specific questions provide the background for subsequent usability-tests that focuses on isolated problems.
- Combined interpretation of standardized scales and specific questions provide the necessary information for strategic decisions that are based on general considerations about the merit of special features.

It is important to note, that benchmarking (no matter if questionnaires or other methods are employed) has to be embedded in the whole process of usability evaluation. We propagate to use a benchmarking questionnaire to establish benchmarks that were measured by a huge sample of (representative) users. We also advice, to make additional focused usability tests with small groups of five to ten people for the identification of concrete usability problems. The next section II summarizes exemplarily a general framework for usability evaluation that illustrates how usability benchmarking can be embedded in the whole process of usability evaluation.

II. OVERALL FRAMEWORK OF USABILITY EVALUATION: USABILITY BENCHMARKING AS ONE IMPORTANT KEY ELEMENT

The proposed model of usability evaluation is based on a combination between systematic quantitative investigations and focussed usability studies. Thereby, both, quantitative as well as qualitative methods are explicitly based

on a systematic empirical approach. Evaluation is conceptualized as research and this is especially true for usability evaluation. However, in praxis, usability evaluation is often done in a pragmatic and unsystematic way. This could sometimes be helpful at very early stages or for transitory applications, but it is not a valid or effective strategy for the establishment of long-lasting Web sites or online services that aim at important goals of public interest (social, medical, educational). Even though the described framework was originally conceptualized for a special application scenario, the usability evaluation of digital libraries [8], it is also applicable for every modern Web site and online services that have to consider new technological possibilities and new requirements of the users.

The proposed multi-method approach is conceptualized as an iterative cyclic process (like it is state of the art in the field). The four core elements are repeated benchmarking, focussed usability studies, derivation of recommendations and decision on the planned improvements in the face of the overall strategy. Repeated benchmarking by a questionnaire delivers quantitative indicators based on a large panel of users. This enables comparisons with prior versions of the Web site (as a quality check of improvements), with competitors as well as with general standards of usability. Additionally, the data allows the prioritization of usability problems and user requirements and can build up the basis for subsequent focused usability studies. These focussed usability studies work mainly with qualitative data and thinking-aloud method. In case, also quantitative indicators (e.g., error rate) can be incorporated. Usability recommendations will be based on the results of usability benchmarking as well as on the data of the focussed usability studies. Normally, it makes sense to interpret both data sources together, to give more specific and accurate recommendations. The decision about the next improvements of the Web site will be made in the light of the usability recommendations as well as in the light of overall strategic decisions. The latter point can be the most difficult one, especially if the content of the Web site is of public interest. Think of a Web site for teenagers: Even though teenager might wish to have more free music downloads and violent video games, this might contradict the policy of the owner of the Web site (for example a Christian organisation).

The four elements are closely interconnected and can be flexible combined. More details are given in [8].

III. CREATING A QUESTIONNAIRE FOR USABILITY BENCHMARKING

A. Advantages and Limitations of a Benchmarking Questionnaire

Like mentioned before, a benchmarking questionnaire has several advantages. First of all, a questionnaire has an uncomplicated handling. Multiple choice questions and scales can be quickly be filled out and the data recoding and analysis is relatively easy and unambiguous. (Contrariwise, the analysis of behavioral data can be very time- and resource-intensive.) That means, a questionnaire is often less resource-intensive and time-consuming (compared to behavioral measurements).

For the assessment of usability, there are several standardized questionnaires existing, short ones (e.g., SUS by [9]) as well as more detailed ones (e.g., ISONORM by [10]). Standardized scales enable (at least partly) the comparison with general standards an, competitors and prior versions of a web site in a valid and reliable way. However, besides the known limitations of subjective data, existing standardized questionnaires are sometimes not adequate for the specific Web site and its services. Thus, the usual standardized usability questionnaires have to be complemented by specific questions. Additionally, questionnaires normally concentrate on a judgement of the existing features, but it is not assessed, what is missing. This means, the quantitative benchmarking process should be also enriched by supplementary questions that deliver qualitative prospective data.

The questions and items of scales should be formulated in a distinct and unambiguous way in order to assure, that the assessed data comprise meaningful information, which can be easily interpreted. (Contrariwise, it is often complicated to interpret the meaning of behavioral observations like starring at a Web site: Why is the user staring, what is going on in his/her head? Fascination? Irritation? Confusion? Is she/he bored or sleepy?).

Since no moderator is necessary, it can be easily distributed over a big sample of people. In some situations it is necessary to have a moderator, but normally this can be managed in a group session, i.e. more than one person can be tested at one time. General advantages and drawbacks of subjective self-reporting data versus behavioral data are disputed in [10].

A few additional annotations regarding the presentation of a questionnaire: Normally, no moderator is necessary. The questionnaire can be presented paper-based or on the computer either online or offline. It can be handed out in a face-to-face situation or in case can also be sent out by mail or email. In case, the benchmarking questionnaire can be also presented as internet survey. However, it is important to note, that these different forms of data assessment are connected with different levels of control. In a face-to-face situation you have the highest control about the situation and the procedure. Even though a moderator is not necessary for a questionnaire it can be advantageously. Sometimes, people have questions or they are insecure, if they understood the scale in the right way. Additionally, a moderator can make notes, if a question is especially hard or if a specific person needs much more time than the other participants. Contrariwise, in the absence of a moderator, no one knows what people eventually do in parallel while they filled out the questionnaire: Watching TV, surfing in the internet, talking with other people about the questionnaire. Especially the latter point bears the danger to receive a biased answer. This situation gets worse, when the questionnaire is presented as internet survey. In principle, an internet survey produces a selective user-sample: Only people who visit the page will participate. Thus, one has to ask which user groups normally visit the page: is this an adequate, representative sample? Additionally, it is hard to control, if the data are valid: was the participant actually – like claimed in the personal data – a rich old lady or rather a young guy with too much spare time. Another danger is, that people can participate a second or third time, for example because they want to have the reward for a second time.

Rewards are another important point (not only for questionnaires, but for every study with participants). If your participants are volunteers, this will save a lot of money,

but on the other hand it can also bias the data. So you have to ask why people participate voluntary. Why do they sacrifice their spare time for your study – instead of going out and eating ice cream in the sun? Volunteers have a often a special motivation. Maybe they participate for free, because they like the product or Web page (or they really hate it). This motivations bias the data and thereby you have a selective sample – but not a representative one.

Like mentioned in section I, it does not always make sense to have a benchmarking questionnaire. If you are in the very early phases of website development and have only some prototypes and wireframes it makes obviously no sense to set a benchmark. Additionally, if you want to eliminate very specific usability problems, a benchmarking questionnaire can give you only a rough impression, what is going on. The potential concrete obstacles in the software or online service will be better tested by a usability-test. As described in section II the process of usability evaluation comprises more than only a benchmarking questionnaire. The combination of the different elements and methods has to be designed in the face of the concrete demands. However, by means of specific questions of a benchmarking questionnaire one can identify the appropriate context and scenario / part of the Web site which is problematic and should be tested in subsequent usability tests.

B. Benchmarking Measurements in the Light of Usability Goals

The goals of usability evaluation and the used benchmarking questionnaire could be manifold. The following list gives an overview of the most important aspects:

- Addressing general or isolated usability questions including the identification of general and specific usability problems of an existing Web site
- Comparison with competitors or general standards of usability (if available)
- Quality check and quantitative measurement of improvements and the merit of innovations
- Formulation of concrete usability recommendations for the future
- Assessment of user requirements

The list illustrates the broad rage of possible usability goals. Accordingly, the used questionnaire (or other methods of usability assessment) could be rather divers. Thus, in the light of available resources and the required input, a benchmarking questionnaire could be either short or relatively long. The used scales and questions have to be carefully designed with respect to the concrete usability goal(s) and the prospective planned developments of the Web site. In the following, we provide a general overview on the possible elements and composition of a benchmarking questionnaire.

C. Quantitative and Qualitative Measurements

Taking the considerations above, it is obvious, that quantitative measurements are the very heart of a benchmarking questionnaire. However, this doesn't mean, that there are no qualitative measurements possible. One might add some open question to receive qualitative information regarding user requirements. For example, in the ISONORM [10], after the ratings of a specific criteria (e.g. error tolerance), it is asked for a concrete example, that illustrates the users experiences. These individual examples can deliver fruitful qualitative input for usability recommendations. Additionally, one might add a labeling test or presents some screenshots or alternative paper-prototypes and ask for open feedback. These qualitative data support the interpretation of the quantitative ratings or give first evidence on the user's opinion regarding planned innovation. This in turn provides also the basis for the next benchmarking cycle (i.e. if the questionnaire has to be modified by adding new scales or further concrete questions).

D. Using Standardized Scales

We propagate to use at least one standardized scale because a standardized scale comprises crucial advantages:

- The items are already tested by many people: Thus, it can be assumed, that the wording is comprehensible
- Reliability (see e.g. [11]) and validity (see e.g. [12]) are tested
- Standardized scales are often applicable for a broad range of variations of the same thing. Thus, even after a substantial change of the web site or application the scale can be used without changing the items. Accordingly, a direct comparison between different versions of a Web site is possible. This is essential for a quality check. Changes are not necessarily a merit.
- General standards are partly available. For example, for the SUS also the concrete meaning of the values is tested [13]. This means, based on your assessed values one can decide if the own Web site / application comprises a low, sufficient or excellent level of usability.
- Sometimes also the indices of leading web sites /products or key populations are available. This enables the judgment, if the own Web site is below or above average. Additionally, in case also a direct comparison with competitors is possible.

Examples of popular standardized scales for the assessment of usability are the SUS [9], ISONORM [10], and IsoMetrics [14]. While SUS is rather short, the ISONORM and IsoMetrics are rather long, more detailed instruments. When selecting one (or more) of the standardized scales, one has to keep in mind, what he/she wants to know: General global rating or specific rating of single aspects? Additionally, it is also a matter of practicability, e.g. the available resources for giving the participants rewards (in dependence of the length of the questionnaire) and the need for the inclusion of other scales and methods.

Besides standardized scales for the assessment of usability also other already existing measurement instruments might be helpful. Imaging the usability evaluation of a serious game: In this case the game designers are normally not only interested in the handling of and navigation through the game, but also in the amount of game-play, the flow-experience and the amount of learning. Even though these aspects are not part of usability in a narrow sense, from a broader perspective the game-play and flow-experience are part of the "joy of use" and the learning experience is part of an "effective and efficient" use (see definition of usability in the section I).

E. Using Specific Questions and Specialized Internal Scales

Many Web sites and online services are very specific. Also usability goals can be very specific and unique for a company or owner of a Web site. In such cases, we need specific questions to assess the required information. Such specific questions might address requirements of the users, planned innovations, the quality of a service (e.g. quality of search results), or special functionalities (e.g., joy of use because of having different emoticons available). In some cases it could be also necessary to develop a scale.

For the construction and wording of the specific items one should discuss the issue with the accordingly content experts. For example, if one wants to assess the quality of a cooking recipe, one has to talk with the cook or owner of the recipe, what are the quality criteria: should it be spicy or wealthy or both? Should it be something that can be made quick and easy or is it more important to have a nice look? Analogously, for the quality of search results, one has to talk with the content experts what are important quality criteria of results (e.g., scientific value, recent findings, exhaustive list, peer-reviewed journals).

The format of the additional specific questions should be designed in a way that makes quantification distinct and easy, i.e. avoiding open answer format as far as possible. Rating-scales, multiple choice or simple yes-no options are advantageously. However, depending on the concrete demands of usability evaluation, also selected open questions can be added to receive necessary input for planning specific usability studies. For example, an open question for the normal usage of the digital library could be very helpful to create a use case for a subsequent usability test with single end-users.

If the construction of a new scale is necessary, we propagate to construct it in the analogous format like the used standardized scale. For the participants it is much more comfortable and less confusing to manage an unique answering format and not switching between different rating-scales. For example, if one uses the SUS with its 5-point rating-scale (from strongly disagree to strongly agree), then the additional scale should be designed with the analogous answering format. (Of course, the answering format of existing standardized instruments should never be changed. Even though using two or more standardized questionnaires with different scaling, leave them as they are. Changing the answering format destroys most advantages of standardized questionnaires, namely reliability, validity and the possibility of comparison with existing benchmarks and general standards.)

However, in some cases it makes sense to use a different answering format, e.g. if you want to underline, that the scale is on something completely different or if you need percent-ratings. Sometimes, for single ratings on additional questions, a rating scale from 0 to 10 is the most intuitive answering format.

In all cases, it is important to make some pilot tests with newly constructed items, i.e. test the items and additional questions with three to five people. Ask them, how they understand the wording and check if this is in line with the intention of the question. Please keep in mind: The question shapes the answer [15] [16].

F. Additional Methods

So far we have disputed questionnaires with scales and open questions. However, in some cases it could be advantageously to enrich a benchmarking questionnaire with additional methods. Examples are: scribbling in a screenshot, completion of a design with hand drawing, or putting labels on a rough sketch. For a concrete feedback, sometimes it is much easier to write it directly in the screenshot instead of giving verbal comments (Think of comments like: “It confuses me that the little button in yellow in the second half of the page has a similar label compared to the orange one at the right top.”, “I would arrange the buttons A, B and F in one line and putting the pictures that are now on the right half on the page rather on the left bottom.”). Especially, if one uses many graphics and visualizations or wants feedback on the arrangement of elements on a web page this can be advantageously. Normally, the focus of such methods is on qualitative data. However, the input can also serve as a quantitative indicator: For example, how many negative or positive comments were made or how many people marked a button as confusing. Even though it is rather resource-intensive to analyze such data in a quantitative way (compared to rating-scales), it might be helpful for some design decisions.

G. Control Variables

Control variables are very important for data interpretation. Examples of usual control variables (not only in the context of Web usability) are:

- Gender, age and profession
- Level of experience with the computer, the internet and/or the specific online services
- Motivation for participating the study, personal goals (children, money)

The according variables can give additional insights to the data and might reveal important differences in your user group. Imagine that you have incorporated an innovative feature using 3D-visualizations. In sum, you might find no differences in the usability evaluation between the version with and the version without this innovation. However, if you integrate the control variable gender, you could find that females rated the usability of the 3D-innovation rather low whereas males judged the usability rather high. Contrariwise the use of emoticons could be very welcomed by your female users but might be judged as needless by your male users.

To sum up, control variables provide additional information that might lead to a refinement of usability benchmarks as well as to a refinement of usability recommendations.

H. Appropriate Selection, Combination and Order of the Different Elements

So far we have described different possible elements of a benchmark questionnaire. For the decision which of them should be included in the benchmarking questionnaire for the concrete application scenario, one has to analyze the concrete goals of the usability evaluation.

In principle, a benchmarking questionnaire comprises the following main elements:

- Welcome and introduction of the purpose of the questionnaire as well as instructions, how to handle the questions and scales

- At least one standardized existing scale for the general assessment of usability (short or long one)
- Specialized scale or specific questions with respect to the most important functionalities or features that are in the centre of the usability evaluation
- Control variables (age, gender, level of experience etc.)
- Some space for additional open remarks

For a comparison with the baseline-benchmark, these elements have to be kept constant in the subsequent cycle of usability evaluation. That means if you decided to use the SUS as standardized global scale, you should use it for every usability evaluation cycle. Analogously, also the specialized questions and additional internal scales have to be kept constant as far as possible. This again underlines the importance of pilot tests: The understandability of the wording of the items has to be tested in advance. If you are forced to change the wording after the benchmarking process (because you have recognized, that half of the people didn't answer the items but making question marks besides the scale) it is too late and you cannot use the data for benchmarking or comparison with the next version. (For two reasons: First, the assessed data are not valid, because the wording was not clear and you have many drop-outs. Second, after you have changed the wording the next data sample is not comparable, because you have other items.)

Besides these main elements of a benchmarking questionnaire, you might include some further elements. Especially, if innovations or changes are planned, it is advantageously to add some further questions that are not part of the benchmarking process, but help to concretize usability improvements or support the interpretation of those usability ratings which result from the standardized scales. It depends of the available resources and the concrete goals of the usability evaluation process if and which of these questions make sense:

- Prospective questions on planned innovations
- Questions on user requirements
- Assessment of the liked and disliked features; open questions on obstacles and welcomed elements

The results of these additional questions are not used for benchmarks. Rather they serve to add surplus meaning to the benchmark and help designers and managers for future decision. Thus, these additional questions (normally) have to be modified in the next evaluation cycle.

For the order of the single elements its obvious that welcome and introduction as well as the instructions have to be in the very beginning. Control variables can be assessed either before or after the usability scales and questions. If you have scales for different elements of features, the single scales and questions should be presented in different orders to make sure, that the order of the question doesn't influence the answer. (There is a long research tradition on order effects in questionnaires, for details see e.g. [17], [18], and [19]. There are not always order effects – but you'll never know until you have tested it. Thus, having the questions in different orders is the safer way.)

When selecting the scales and questions it has to be kept in mind, that the participants should not be overburdened. If the participation is voluntary, the patience is often much lower than in case of payment. (The differences between voluntary participation and rewards were already discussed.) But in any case: Test the time that is needed for completing the questionnaire. Make pilot tests, eventually shorten the questionnaire or use a shorter scale instead of a longer one. Concentrate on the most important issues. Of course, there are always some more interesting aspects, however keep in mind what's the core: Usability benchmarking. If the questionnaire is too long and participants are tired or annoyed, you will not receive valid usability data, but rather a feedback on the frustrating survey.

IV. DESCRIPTION OF A CONCRETE EXAMPLE

In the following, the described general methodological considerations will be explained by the concrete example of the digital library services of the ZBW – Leibniz Information Centre for Economics (http://www.zbw.eu/index-e.html). The ZBW is the world's largest specialist library for economics, with locations in Kiel and Hamburg. The ZBW provides numerous online services like EconBiz for literature search or EconStor for publishing working papers. But until now, the usability of the online services and the Web site itself has not been evaluated in a systematic way. Furthermore, the services will be enhanced and improved in the future. In the rise of Web 2.0 technologies also digital libraries and information centers have to face new challenges. While most modern digital libraries have nowadays electronic resources, search engines and divers online services, so far the integration of the Web 2.0 technologies is rather spare or incomplete. Furthermore, the new generation users (not only of digital libraries) have new requirements. These requirements address not only Web 2.0 technologies but also a sufficient level of usability. Usability is conzeptualized as a key factor to attract users that would otherwise use Google scholar or similar search tools.

Therefore, an internal task force for usability evaluation was established at the ZBW. Accordingly, the starting point of the usability evaluation at the ZBW was the strategic decision to become a modern Web 2.0 library with a good level of usability. To reach this strategic aim, a redesign of the Web Site and continuous improvements of the services are planned. Both, the re-design as well as the improvements of the services have to be evaluated, i.e. it has to be assured that the changes are perceived as improvements and the level of improvements should be controllable and quantifiable.

It's important to note that in this scenario, an existing online platform build up the starting point. The upgrades, enrichments and extensions of the platform are in the center of usability evaluation. Thus, it is essential, that changed features can be directly compared with prior versions or in other words: The changes have to be made quantifiable. Thus, in a first step, a benchmarking questionnaire was created for a baseline assessment of the status quo as comparison standard. Thereby, four main objects are in the center of usability evaluation: the homepage of the ZBW, the online service for literature search (called EconBiz), the publishing portal (called EconStor) and the expert online help (called EconDesk). These four objects build up the very heart of the ZBW online and thus are also in the center of the planned improvements and the accordingly usability evaluation. The

benchmark questionnaire comprises standardized existing usability scales for the named four evaluation objects, an additional scale for assessing the quality of literature search results, and several prospective questions to receive additional qualitative data.

The benchmarking questionnaire serves different usability goals. First, to measure the status quo as a baseline that can be used as benchmark and comparison-standard for quality check. Second, to identify the most important usability problems – that should be avoided in the redesign. Third, to assess user requirements and to receive feedback regarding the planned innovations.

In order to address these multiple goals, the benchmarking questionnaire for the ZBW is rather long. It consists of several elements that could be partly changed or omitted in the next benchmarking cycle. The following list gives an overview of the components:

- Introduction and instructions
- Assessment of selected personal data and control variables
- Standardized scale SUS for the ZBW homepage, EconBiz, EconStor, and EconDesk
- Standardized scale ISONORM for EconBiz
- Internal newly constructed Scale on the Quality of Literatur Lists (SQuaLL) for assessing the quality of literature search
- Additional open questions to assess user requirements and first feedback on the planned innovations/changes
- Advanced scribbling for informal feedback as important input for the planned redesign of the Web Site

The single elements will be described in the following subsections. Afterwards the exploitation of the data and an outlook for possible modifications will be given.

A. Introduction, Instructions, Personal Data and Control Variables

The questionnaire starts with a short introduction that explains the purpose of the questionnaire and named the responsible person and the contact person for privacy issues. Thereby, it is made explicit, that the usability-department works as an independent task force. This should ensure that politeness effects are avoided.

While most standardized questionnaires (see below) have their own standardized instruction, for a composed long questionnaire it makes sense to present also some general information in the beginning. One very important point is to make clear, that there are no right or wrong answers. The questions address the individual personal opinion of the participants and the participants were only asked to give open and honest answers.

Since the questionnaire is rather long, participants receive a 20,-€ voucher (for a popular internet shop) as reward for participation. During the study, refreshments and some sweets are offered.

The questionnaires are filled out (offline, paper-based) in the presence of a moderator in group sessions. Participants can ask during the whole study for help. In principle, the questionnaire could be also presented online or sent out by mail. However, we choose this more controlled setting for two reasons: First, we ask all participants if they are willing to participate also in other usability studies and thus, we want to get to know or even see the person for the first time (remember the possible misuse of internet-surveys). Second, the data of the baseline-assessment are also used for a research question on order-effects (namely part-whole effect) and thus, a controlled setting is needed for scientific reasons.

As control variables, the age, gender, family status, profession, experience with the internet, computer use as well as the experience with the ZBW online is assessed. These control variables enable not only the identification of different subpopulation of users but also allows identifying the influence of prior experiences (with the computer, the internet or the ZBW online) on the usability estimation.

B. Standardized Scales

We selected the SUS [9] as a short standardized measurement of the usability of the four main objects of usability evaluation: the ZBW homepage, the online service for literature search (called EconBiz), the publishing portal (called EconStor) and the expert online help (called EconDesk). These global ratings are the core of the benchmarking process and the assessment will be repeated for every subsequent cycle of usability evaluation (i.e. after the redesign has been finished and after other substantial changes in the future).

For the core service of the ZBW, the online literature search EconBiz, also a more detailed standardized usability scale, the ISONORM [10], is included. The ISONORM is rather long and thus, it is not possible to use it for all the four evaluation objects, because this would overburden the participants. Additionally, the questions are formulated in a very detailed way that is partly not suitable for EconDesk and EconStor. However, for the literature search service such detailed information is needed, because EconBiz has several sophisticated functionalities. From prior feedback from the ZBW-users it was known, that these functionalities are partly too complicated and partly unfamiliar. Thus, we want to have structured systematic data on this service that enable a concrete improvement of usability and provide the basis for subsequent usability tests.

C. Specific Internal Scale on the Quality of Literature Search

Like pointed out above, the online service EconBiz has many sophisticated functionalities, and therefore offers access to a high quality literature. Roughly spoken: The literature search with EconBiz is quite demanding for the user (compared to tools like Google), but therefore has the merit of a much higher quality of the search results. Thus, for an overall evaluation of EconBiz, both aspects have to be assessed: the handling and the merit for the users. The general usability is assessed by SUS and ISONORM, but ratings on content quality are not included in these standardized usability scales. (It was also a desire of the ZBW management and the product managers of EconBiz to evaluate the quality of the results of the literature search.) In a broader sense the content quality can be seen as a part of a detailed usability assessment (user satisfaction).

For the assessment of the quality of the literature search results an own internal scale was constructed. The format of the newly constructed **S**cale on the **Qua**lity of **L**iteratur **L**ists (SQuaLL) was designed analogous to the SUS.

Therefore, we first made an expert query by means of a semi-structured interview with internal experts of the ZBW. Based on the expert interview, ten important quality criteria of literature search were extracted. (Of course, other or additional quality criteria are possible. The quality criteria of the SQuaLL were selected in the face of the specific demands of the ZBW as a specialized information center for economics). For each of the ten quality criteria an item was constructed. The items are formulated in a way that is analogously to the SUS. For the rating of the items, the same 5-point rating-scale as the SUS (from strongly disagree to strongly agree) is provided. Table I shows the quality criteria and the accordingly items.

Even though neither validity nor reliability are proved so far, the data deliver a useful internal benchmark for the most important quality criteria of EconBiz. The benchmark data can be used as comparison standard for improvements or enhancements of EconBiz. In the future, the scale will be tested for (parallel-test) reliability and (criteria-) validity.

To sum up, the SQuaLL is a good example, how standardized scales (like SUS and ISONORM) can be complemented by specialized items and scales. To judge the usability of EconBiz in a more holistic sense, both are needed: A general usability rating of the handling as well as an estimation of the effectiveness and appropriateness for the subjective aims in the course of a literature search.

TABLE I.
ITEMS OF THE SQUALL

SQuaLL Items	Description	
	Quality criteria	*Wording of the items*[a]
1	Exhaustiveness	The result-list of EconBiz provides me an exhaustive overview on the subject area.
2	Reasonable ranking	The ranking of the result list of EconBiz well elaborated.
3	Scientific proofed quality	I can be sure, that the listed contents of EconBiz are of high scientific quality.
4	Relevance (based on meta data)	The listed results in EconBiz are in very good accordance with the used search keys.
5	Valid and modifiable filters	The filter functionalities of EconBiz are flexible modifiable and select relevant items form irrelevant issues in a reliable way.
6	Events included	I estimate the listing of events within the literature list as very informative and interesting.
7	Availability of full texts is visible	From the result list of EconBiz I can immediately recognize, if and where a full text of the reference is available.
8	Export functionality	I can export the listed references without any problems, e.g. integrate it in my own literature list or software tools for managing references.
9	Additional information	For the single references of the result list of EconBiz there is helpful additional information given.
10	Information on new releases available	Via EconBiz I can also look for new releases.

a. Original items are in German. Interested readers can contact the first author for the German version.

D. Advanced Scribbling as Qualitative Data Source

To receive an open and less informal feedback on the existing homepage and online services, a playful scribbling task was created. Users receive a screenshot of the existing homepage and the services EconBiz, EconDesk and EconStor. For the service EconBiz we use two screenshots, one for the start-page and one for the list of search results, because they are very divers in their design, content and functionalities. Additionally, a blue pen as well as a green, a red and a yellow text marker are provided. Thereby, colors follow the analogy of a traffic control light: Red should be used for elements that are perceived as confusing or not understandable, yellow for elements that are needless, and green for elements that are important for the user. The users are instructed to mark on the screenshot the respective elements with red, yellow and green. Additionally, the participants are free to add comments, annotations, and critique with the blue pen. (In the pilot tests this task was really fun for the participants: they like the painting and scribbling with different colors for providing concrete and direct feedback on the pages and their layout. Thus we put this task in the middle of the long questionnaire, to give the participants a kind of cognitive relaxing break.) The meaning of the colors is not only explained in the instructions but is also visible as legend at the right bottom besides the screenshot.

Even though this scribbling task delivers mainly qualitative data, one can also make quantitative use of it, for example: Calculating, how many areas are marked with red, calculating which areas are marked green by most of the users, or calculating, how many negative or positive annotations are given. In principle, also these quantifications can serve as a benchmark (e.g., ratio of negative and positive comments). However, since a quantitative analysis of the scribbling task would be more resource-intensive and less objective than the analysis of the questionnaire data, we use it only in a qualitative way to provide the product designers the necessary input for the required changes in the layout of the pages.

E. Additional Specific Questions: User Requirements and Prospective Issues

In the light of the planned changes and strategic decisions, several additional questions are of interest. These questions addresses aspects of Web 2.0., personalization, advanced functionalities of the literature search and competitive services for literature search (including Google and Google scholar). The ZBW has been started to use Web 2.0 applications and it is planned to extend these activities. As a modern library 2.0 the ZBW is especially interested in the user requirements of these innovations. Thus, we incorporate questions on the general use of the ZBW and the internet as well as questions on the wish to communicate with other ZBW users. Background of these questions is the ongoing discussion how and to which extend a personalization of user account and an internal platform for professional online discussion is desired.

Additionally, several open questions on the prior experiences with EconBiz, EconStor and EconDesk are included. These open questions are formulated in collaboration with the responsible product managers, i.e. the data serve directly to answer open questions of the product-developers. At the very end of the questionnaire, there are also four very general open questions, that give participant the chance to provide feedback about their own urgent

issues (which might have been overlooked). These four open questions are: What do you like at the ZBW? What do you dislike at the ZBW? What do you wish for the future at the ZBW? Additional recommendations and critique.

F. Outlook: Changes for the Next Benchmarking Cycle

Like explained above, the single elements of the baseline questionnaire serves different usability goals. Important keystones are the SUS for the homepage, for EconBiz, for EconStor and for EconDesk as well as the additional internal scale SQuaLL. These measurements will be presented in each benchmarking cycle. The presentation of the other elements is dependent on the future developments at the ZBW. Based on the concrete demands in the next cycles, a modified questionnaire will be presented.

The next benchmarking cycle after the baseline assessment will take place after the complete Web site has been re-designed and bugs were consolidated. In between focused usability tests will be conducted to identify concrete usability problems. The derived usability recommendations have to be aligned with the strategy of a modern library 2.0 and the available resources (and privacy issues). Accordingly, also external associated links (ZBW on Facebook and Twitter) might come into play.

V. RESUME

Usability benchmarking is an important keystone in the process of usability evaluation since it provides a standard for comparison and enables the quantification of improvements. The theoretical explanations pointed out that there are different possibilities how a benchmarking questionnaire could be constructed and that specialized questions and scales have to be designed in relationship with the concrete usability goals.

The practical example of the benchmarking questionnaire of the ZBW illustrates also how the basic methodology and wide-spread methods and scales can be enriched by more creative approaches that are designed for the specific needs of the concrete Web site. The combination of the SUS for EconBiz and the Specialized Scale for the Quality of Literature Lists (SQuaLL) underpine, how the combination of standardized scales and specific question can enable a more holistic and appropriate assessment of the usability of a very specialized service.

Generally, benchmarking questionnaires are not standalones, but should be embedded in a cyclic, repeated process of usability evaluation (following state-of-the-art methodology). Quantitative and qualitative methods complement each others. Additionally, subjective data could (and should) be enriched by objective measurements (behavior observation, recording of logfiles etc.).

In the end, besides every measurements and every creative design idea, the most important thing for ensuring a good usability is neither content nor design, but rather the use of it.

REFERENCES

[1] J. Rubin and D. Chisnell, Handbook of usability testing. Indianapolis: Wiley Publishing Inc, 2008.

[2] International Standard Organization, “Ergonomic requirements for office work with visual display terminals. Part 11: Guidance on usability (ISO DIS 9241-11),” London: International Standards Organization, 1994.

[3] J. Nielsen, “Usability engineering,” Cambridge, MA: Academic Press, 1993.

[4] T. Brinck, D. Gergle and S. D. Wood, “Designing Web sites that work: Usability for the Web,” San Francisco: Morgan Kaufmann, 2002.

[5] P. Booth, “An introduction to human-computer interaction,” London: Lawrence Erlbaum Associates, 1989.

[6] N. Bevan, “International standards for HCI and usability” International Journal of Human-Computer Studies, vol. 55, pp. 533-552, 2001. http://dx.doi.org/10.1006/ijhc.2001.0483

[7] S. B. Linek, B. Marte and D. Albert, “The differential use and effective combination of questionnaires and logfiles,“ in Computer-based Knowledge & Skill Assessment and Feedback in Learning settings (CAF), Proceedings of the International Conference on Interactive Computer Aided Learning (ICL), 24th to 26th September, 2008, Villach, Austria.

[8] S. B. Linek and K. Tochtermann, “Sophistikated usability evaluation of digital libraries,” in Proceedings of the 10th European Conference on e-Learning (ECEL 2011), 10th to 11th November, 2011, Brighton, UK.

[9] J. Brooke, “SUS: a „quick and dirty“ usability scale,” in Usability Evaluation in Industry, P. W. Jordan, B. Thomas, B.A. Weerdmeester and a. L. McClelland, Eds. London: Taylor & Francis, 1996. Retrieved March 24th 2011 from: http://hell.meiert.org/core/pdf/sus.pdf.

[10] J. Prümper, “Test IT: ISONORM 9241/10,” in Human-Computer Interaction - Communication, Cooperation, and Application Design, H. J. Bullinger and J. Ziegler, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1999, pp. 1028-1032.

[11] L. S. Feldt and R. L. Brennan, “Reliability,” in Educational Measurement, 3rd Ed., R. L. Linn, Eds. New York/London, 1989, pp. 105-146.

[12] S. J. Messik, “Test validity and the ethics of assessment,” American Psychologist, vol. 35, pp. 1012-1027, 1980. http://dx.doi.org/10.1037/0003-066X.35.11.1012

[13] A. Bangor, P. Kortum and J. Miller, “Determining what individual SUS score mean: Adding and adjective rating scale,” Journal of Usability Studies, vol. 4(3), pp. 114-123, 2009.

[14] G. Gediga, K. C. Hamborg and I. Düntsch, “The IsoMetrics usability inventory. An operationalisation of ISO 9341-10 supporting summative and formative evaluation of software systems,” Behaviour and Information Technology, vol. 18, pp. 151-164, 1999. http://dx.doi.org/10.1080/014492999119057

[15] N. Schwarz, “How the questions shape the answers,” American Psychologist, vol. 54(2), pp. 93-105, 1999. http://dx.doi.org/10.1037/0003-066X.54.2.93

[16] N. C. Schaeffer and S. Presser, “The Science of asking questions,” Annual Review of Sociology, vol. 29, pp. 65-88, 2003. http://dx.doi.org/10.1146/annurev.soc.29.110702.110112

[17] N. Schwarz, F. Strack and H.-P. Mai, “Assimilation and contrast effects in part-wholequestion sequences: A conversational logic analysis,” Public Opinion Quaterly, vol. 55, pp. 3-23, 1991. http://dx.doi.org/10.1086/269239

[18] F. K. Willits and B. Ke, “Part-whole question order effects. Views of rurality,” Public Opinion Quaterly, vol. 59, pp. 392-403, 1995. http://dx.doi.org/10.1086/269483

[19] C. W. DeMoranville and C. C. Bienstock, “Question order effects in measuring service quality,” International Journal of Research in Marketing, vol. 20, pp. 217-231, 2003. http://dx.doi.org/10.1016/S0167-8116(03)00034-X

AUTHORS

S. B. Linek and **K. Tochtermann** are with the Leibniz Information Center of Economics, Düsternbrooker Weg 120, 24105 Kiel, Germany (e-mails: s.linek@zbw.eu and K.Tochtermann@zbw.eu).

This article is an extended version of a paper presented at the International Conference ICL 2011, held in September 2011 in Piešťany, Slovakia. Submitted, October 30th, 2011. Received 28 September 2011. Published as resubmitted by the authors 22 November 2011.