Linek, Stephanie

Conference Paper

Usability evaluation of digital libraries: offline and online assessment compared


This Version is available at:
http://hdl.handle.net/11108/195

Kontakt/Contact
ZBW – Leibniz-Informationszentrum Wirtschaft/Leibniz Information Centre for Economics
Düsternbrooker Weg 120
24105 Kiel (Germany)
E-Mail: info@zbw.eu
http://zbw.eu/de/ueber-uns/profil/veroeffentlichungen-zbw/

Terms of use:
This document may be saved and copied for your personal and scholarly purposes. You are not to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. If the document is made available under a Creative Commons Licence you may exercise further usage rights as specified in the licence.

Standard-Nutzungsbedingungen:
USABILITY EVALUATION OF DIGITAL LIBRARIES: OFFLINE AND ONLINE ASSESSMENT COMPARED

Stephanie B. Linek
ZBW – Leibniz Information Centre for Economics (Germany)

Abstract
Usability is of special importance for every online product. Also a modern digital library or Library 2.0 [1] needs to assure an easy handling and the usability of a library’s website with the integrated online services have to be continuously monitored. An efficient way for usability assessment is the use of short standardized scales either offline (paper pencil, face to face) or online (e.g., by an Internet survey). Offline and online assessment have their special advantages and drawbacks. Thereby, the differences between offline versus online assessment might cause different data even though using the same scales (so-called mode effects). Thus, it has to be assured that the offline measurement delivers equal results like the analogous online measurement.

In this contribution I present an empirical investigation on potential mode effects for a questionnaire for the evaluation of a Library 2.0. The concrete use case of a Library 2.0 was the ZBW (www.zbw.eu), the world’s largest information center for economic literature. The questionnaire was applied in an offline versus an online version. Both versions regard to the evaluation of the homepage of the ZBW as well as the three core services of the ZBW (EconBiz for literature search, EconDesk for online help, and EconStor, a publishing portal). As measurement instrument I used the System Usability Scale (SUS, see [2]) as a short standardized usability scale for each, the homepage as well as the three services. In addition to the usability evaluation, for the service EconBiz also a quality evaluation by the standardized short Scale on the Quality of Literature Lists (SQuaLL, see [3]) was implemented. The SUS and the SQuaLL were applied for the offline version and the online version analogously, i.e., with the same wording and the same answering options. This enables a direct comparison of the quantitative data (offline versus online) assessed by the SUS and the SQuaLL. Additionally, both versions comprised the possibility to give open comments. Besides these common elements there were also some differences between the two versions: The offline version was a rather long version because it included also additional questions for heuristic insights and more detailed feedback about the ZBW. The online version was a short condensed version of the questionnaire that comprised the main standardized short scales (SUS, SQuaLL), only. The offline version was filled out by paper-pencil in the presence of an instructor (face-to-face setting). The online survey was announced via mailing lists, via the homepage of the ZBW and via the Facebook-page of the ZBW. The results showed that there were no significant differences between the offline versus online evaluation by means of the SUS and the SQuaLL. However, most online participants ignored the possibility to give open comments whereas the offline participants gave rather exhaustive annotations.

Overall, the findings suggest that for a pure evaluation control check a short online survey is sufficient and leads to equal results like a long, more controlled offline questionnaire. However, if deeper insights are needed, an offline evaluation seems to be more appropriate. The theoretical and practical implications of the findings will be discussed.

Keywords: Usability, evaluation, online survey, offline setting, assessment methods, digital libraries.

1 INTRODUCTION AND THEORETICAL BACKGROUND

Nowadays good usability is one essential cornerstone of every successful software product and website. Also public institutions and information centres like a Library 2.0 [1] have to assure an effective and easy handling of their online services. The according usability evaluation and monitoring can be done offline or online. In the following a brief overview on usability evaluation and possible mode effects of online versus online assessment are given.
1.1 Usability Evaluation of a Library 2.0

In principle there are several possibilities how to evaluate the usability of a Library 2.0 and its online services. The selection of the appropriate method has to be done in face of the concrete evaluation target and the required information. Thereby, an iterative continuous usability monitoring is state-of-the-art (e.g., [4], [5]) and a multi-method approach with qualitative and quantitative methods is advantageously (see e.g., [6], [7]). One important cornerstone of such an iterative multi-method approach is the use of questionnaires. Thereby, standardized scales can be easily used for a continuous quantitative monitoring and usability benchmarking of a Library 2.0’s website and its online services [7] [8]. Within a benchmarking questionnaire standardized scales can be also combined with other questions and scales or with more qualitative methods [8]. Accordingly, a benchmarking questionnaire could be relative short or rather long. Additionally, one crucial decision relates to the mode of assessment: offline or online.

1.2 Advantages and Drawbacks of Offline versus Online Assessment

A questionnaire can be presented paper-based or on the computer either online or offline. It can be handed out in a face-to-face setting or can be sent out by mail or email. It is important to note, that these different forms of data assessment are connected with different levels of control. In a face-to-face setting you have the highest control about the situation and the procedure. Even though an instructor is usually not necessary for a questionnaire it can be advantageously. Sometimes, people are doubtful about the wording or they are insecure, if they understood the scale in the right way. Additionally, an instructor can make notes, if one item seems to be especially difficult or if a specific person needs more time than the other participants. Contrariwise, in the absence of an instructor, no one knows what people eventually do (in parallel) while they fill out the questionnaire: Watching TV, surfing in the Internet, or talking with other people about the questionnaire. Especially the latter point bears the danger to receive biased answers. This situation gets worse, when the questionnaire is presented as Internet survey. In principle, an Internet survey produces a selective user sample, because only people who know about the link to questionnaire will participate. Thus, the way of recruitment is crucial to address an adequate, representative sample. Additionally, it is hard to control, if the personal data (e.g., age and gender) are valid. Another danger is that people can participate a second or third time. However, from a practical point of view, a (short) online survey is a very quick and easy way for usability evaluation. Even with very limited resources a repetitive application is possible. Additionally, a much broader group of participants can be recruited because an online survey can be answered independent from time and place. A good more detailed overview on the various advantages and drawbacks of online questionnaires can be found by [9]. Additionally, an offline laboratory setting and the presence of an instructor bear some dangers that normally have no relevance for an online study. For example, sympathy or antipathy for the instructor can influence the answers. Furthermore, the presence of an instructor might hinder shy people to give honest and critical answers. Also the superficial situation of a laboratory setting might cause some data bias.

The exemplarily listed differences between offline versus online assessment make clear that the mode of the assessment (offline versus online) can influence the data (see also [10]). These so-called mode effects are a well-known phenomenon in literature. For example, there are several studies on the comparison of online questionnaires with questionnaires sent out by post. The findings on mode effects are mixed and it seems that it depends on the special field and the concrete questionnaire if the online version and the postal versions are comparable. While some studies found mode effects for the answering time or the length of the answers (e.g., [11], [12], [13]) other investigations found equivalent results for both modes of assessment (e.g., [14], [15], [16], [17]).

In most of the mentioned studies, the online version was compared with the postal version (paper-based, sent out by mail). However, for both modes (online and postal) it remains unclear how the participants answer the questions, because there was no direct control over the assessment situation (see also explanations above). Since the participants were not observed during the assessment situation and there was no instructor (face-to-face) present, there were no concrete information about the circumstances, e.g., if the participants made pauses, if they are distracted while answering or if they are unsure about the handling of a scale. Thus, the presence or absence of an instructor can influence the findings (e.g.,[18]).
1.3 Outline of the Study

This paper describes an empirical study on possible mode effects for a benchmarking questionnaire for the evaluation of a Library 2.0. Thereby, two versions of the questionnaire were compared: one version applied offline (paper-pencil, face-to-face in the presence of an instructor) versus one version applied online (Internet survey, self-administered). The concrete use case of a Library 2.0 was the ZBW (www.zbw.eu) which is the world’s largest information center for economic literature, offline as well as online.

The practical background of the study was an upcoming complete relaunch of the ZBW’s website (www.zbw.eu). As a modern Library 2.0, the ZBW has several online services. The usability of the ZBW’s online services is continuously monitored and systematically improved. (A general description of the usability evaluation of a Library 2.0 can be found by [7]). One essential cornerstone of the iterative usability evaluation of the ZBW was the measurement of the usability baseline before the relaunch. This baseline serves as benchmark for comparison with the usability of the new website and the upcoming improvements (for details see [8]). Accordingly, a benchmarking questionnaire was developed, that assessed the usability of the homepage as well as of the three core services of the ZBW (EconBiz for literature search, EconDesk for online help, and EconStor, a publishing portal). In a first step, a rather long offline benchmarking questionnaire was developed. It included additional questions and scales for more detailed information and for heuristic insights. In a second step a shorter online version was established to enable a quick and less resource-intensive repeated usability monitoring in the future.

The used benchmarking questionnaire was very close to the explanation and exemplarily descriptions given in [8]. The empirical usability evaluation of the former website of the ZBW (until Feb 2014) included the homepage itself as well as the three main integrated services EconBiz, EconDesk and EconStor. The usability evaluation was mainly done by means of the System Usability Scale (SUS, see [2]). Thereby, for the homepage and each of the core services the SUS was applied separately. Since EconBiz for literature search was the core service of the ZBW, also the content quality of the results of EconBiz was evaluated. For this purpose, an additional standardized short scale was integrated, namely the Scale on Quality of Literature List (SQuaLL). The SQuaLL was constructed analogously to the format of the SUS and delivers a global quality index for literature search portals. Details on the SQuaLL can be found in [3].

The basic data assessment for the baseline of the benchmarking process was done by a rather long offline version of the questionnaire. The offline version was presented paper-pencil-based in the presence of an instructor. The long offline questionnaire comprised not only the standardized short scales (SUS and SQuaLL) but also the ISONORM [19] as well as several open questions and a qualitative scribbling-task for heuristic insights. This was a rather laborious kind of benchmarking, but it was needed for two reasons: On the one hand, detailed and heuristic insights for the new design of the ZBW’s website were required. On the other hand, a very controlled setting was needed as a clean starting point for the baseline of the iterative usability monitoring.

For a more economic usability evaluation in the future, a short online version of the baseline-questionnaire was created. The online version comprised only the short standardized scales (SUS and SQuaLL) and a few selected additional questions (control variables and possibility to give open comments). The online assessment took place about one year after the offline measurement. Within this time interval there were no substantial changes of the website (homepage and the three services).

Like discussed above, it is questionable, if the data of a paper-based offline questionnaire in the presence of an instructor are equivalent with the data of a self-administered online questionnaire (even when using the same scales). Thus, the quantitative data of the standardized short scales (SUS, SQuaLL) used in both versions of the questionnaire were statistically compared. Accordingly, the aim of the study was twofold: First, to have a baseline measurement (before the relaunch) of the usability of the ZBW’s website and the three online services; and second, to assure the equivalence of the offline versus the online assessment.

The main research question of this study was: Is the usability evaluation of a long detailed usability measurement in an offline setting (by paper-pencil) equivalent to the usability evaluation of a short online version of the questionnaire?
2 METHODOLOGY

Like already mentioned, the study consisted of two separate parts: In the first part, the usability evaluation was done by the offline version (paper-pencil) in a controlled setting in the presence of an instructor. The second part was done about one year later by means of the online version (Internet survey) in a self-administered manner.

2.1 Research Design and Dependent Variables

The research design of the study was a two group design: offline assessment versus online assessment of the usability evaluation of a Library 2.0, namely the ZBW.

The dependent variables were the usability evaluation by means of the SUS of the ZBW’s website itself as well as the three core services (EconBiz, EconStor and EconDesk). In this paper I use the SQuaLL score as an additional dependent variable, because the SQuaLL was also part of both versions of the questionnaire and thus, can be taken as (additional) quantitative indicator for possible mode effects.

2.2 Description of the Questionnaire: Offline Version and Online Version

2.2.1 Parallel Elements of the Offline and the Online Version

Like described above, there were two variations of the questionnaire, an offline version and an online version. Both version comprised the short evaluation scale SUS for the evaluation of the homepage (homepage-SUS), of the service EconBiz (EconBiz-SUS), of the service EconDesk (EconDesk-SUS) and the service EconStor (EconStor-SUS). Additionally, both versions included also a short quality evaluation of EconBiz by the SQuaLL.

The SUS is a short scale with ten items on different aspects of usability. The items are formulated as statements. The statements can be easily modified with respect to the according object of evaluation (e.g. instead of “software” one could insert “EconBiz” or “the homepage of the ZBW”). Each statement has to be rated on a Likert-scale from 1 to 5. The SUS sum-score is calculated like described in [2]. The possible sum-scores are between 0 and 100. A sum-score of 0 indicates a very poor usability, whereas a sum-score of 100 reflects an excellent usability.

The SQuaLL uses the same format like the SUS. It comprises ten items (formulated as statements) that refer to ten different quality criteria of search portals. The items have to be rated on the same 5-point Likert scale like used by the SUS. The calculated sum-scores are between 0 and 100 whereby 0 indicates a very poor quality, whereas a sum-score of 100 reflects an excellent quality. (The SQuaLL exists in two parallel forms. For the comparison of the offline versus the online assessment I used form A. Details on the SQuaLL and its two parallel forms can be found in [3].)

In addition to the described short scales, the offline version as well as the online version comprised the possibility to give open comments and a set of control variables. The control variables included age, gender, experience with the ZBW as well as self-rated competence with the PC, with the Internet and with information search via the Internet. The self-rated competence was assessed by a 10-point Likert scale with 1 = very low competence and 10 = very high competence.

Remark: Both versions of the original questionnaire were in German. The interested reader can contact the author for a German sample of the offline or online version.

2.2.2 Description of the Offline Version

The offline version was the long version of the benchmarking questionnaire that was applied in an offline setting by a paper-pencil survey in the presence of an instructor. For the main usability evaluation the SUS was used for the evaluation of the homepage (homepage-SUS), the service EconBiz (EconBiz-SUS), the service EconDesk (EconDesk-SUS) and the service EconStor (EconStor-SUS). As short standardized scale for the assessment of the quality of EconBiz, the SQuaLL was applied.

In addition to the described short scales, for the service EconBiz also the rather detailed usability scale IS0NORM and a qualitative “advanced scribbling task” (see [8]) was presented. After the standardized scales and the scribbling task, several questions for additional information and comments on the ZBW were included. Most of these questions were open questions for qualitative
input and improvement recommendations. The control variables were assessed as closed questions. At the end of the questionnaire the participants had the possibility to give further open comments.

(Remark: The results of the ISONORM, the advanced scribbling task and the items solely used for the offline version will not be described here, because these data have no relevance for the comparison of offline versus online assessment.)

2.2.3 Description of the Online Version

The short online version comprised (analogous to the offline version) the SUS for the usability evaluation of the homepage (homepage-SUS) and the three services (EconBiz-SUS, EconDesk-SUS, EconStor-SUS) as well as the SQuaLL. Additionally, some selected additional questions on the ZBW, the homepage and the three services were integrated. At the end of the questionnaire (analogous to the offline version) the described set of control variables was assessed and the participants had the possibility to give open comments.

2.3 Pilot Tests

A first draft of the offline version was tested by a small sample of six people. All six people were employees of the ZBW and familiar with the online services. Their feedback was used to test the comprehensibility and appropriateness of the wording. Based on this pilot test, some slight changes in the wording of questions were made.

Since the online version was an excerpt of the offline version, no special test of the wording was necessary. The implementation of the online version was done by the software SoSci Survey (https://www.soscisurvey.de/index.php?page=home&l=eng). After implementation, internal tests for technical errors (correct filters) and typos were made.

2.4 Recruitment of Participants and Methodological Procedure

The recruitment and methodological procedure were done differently for the offline version and the online version. Thereby it is important to remember that the offline version was longer than the online version, because it comprised some additional questions and a heuristic task (described above). Accordingly, the offline version was connected with a higher expenditure for the participants because of the longer duration of the session and the requirement to come to the building of the library. Thus, also the given reward for participation was higher for the offline sample.

2.4.1 Offline Assessment

The prerequisites for participation in the offline survey were that the participants knew and used the ZBW Homepage as well as the service EconBiz before. It was not an obligation to know also EconDesk and EconStor. (These prerequisites for participation mirrored the majority of the usual ZBW clients.) The participants were recruited personally in the ZBW’s buildings (reading room) in Hamburg and Kiel. Additionally, flyers and mouth-to-mouth recruitment were used.

Altogether 96 participants completed the offline version. As incentive each participant received a 20,- € voucher for a popular online shop. During the completion of the questionnaire, some sweets were offered. The duration of the test sessions varied between a half and one and a half hour.

The questionnaires were answered in group sessions in a separate room in the ZBW buildings in Hamburg or in Kiel. After a short welcome, the participants were informed that the data were threatened anonymously and strictly confidential. Each participant inserted a personal ID on the first page of the questionnaire. This ID was used to mark also the other materials of the survey, e.g. the post-interviews. The offline questionnaire was handled out and answered in a written paper-pencil based format. During the whole study an instructor was present.

2.4.2 Online Assessment

The online survey was conducted about one year after the application of offline version. During this one-year-break there were no substantial changes of the website, neither the homepage nor the three services. Prerequisites for participation (analogous to the offline survey) were that the participants knew and used the ZBW homepage as well as the service EconBiz before.

The online sample was recruited in different ways:
• Via existent mailing lists of ZBW’s users: As incentive the participants received a 15,- € voucher for a popular online shop.

• Additionally, there was a New Year’s Eve lottery for recruiting additional participants. The lottery was announced on the homepage of the ZBW and the ZBW’s Facebook-page. Everyone who took part in the lottery could win one of thirty vouchers about 30,- € for a popular online shop.

The link to the online version was sent out by email or was embedded in the announcement of the lottery. There was no back-button in the online survey, i.e., the participants couldn’t change their answers after they completed a scale. For each scale there were forced answers, i.e., only if all items were answered, the participants could go forward to the next questions. At the end of the survey they were asked for their email address (for the participation in the lottery) and a personally created ID (for internal control). In order to guarantee anonymity these indications were voluntary and were handled separately from the data.

3 RESULTS

3.1 Description of the Sample

Altogether, 189 subjects participated in the study. All participants were familiar with the ZBW’s homepage and EconBiz (because this was the prerequisite for participation). The other two online services, EconDesk and EconStor, were partly unknown. The offline subsample comprised an equal amount of male and female participants. The online subsample contained more male participants. The offline and the online subsample were comparable in age, experience with the ZBW and self-rated competence with PC (on a 10-point rating scale), with the Internet, and with information search via the Internet. Details on the two subsamples (offline versus online) are provided in the following.

3.1.1 Offline Subsample

The offline subsample contained 96 (47 male and 49 female) participants. The age of the participants was at average 27.33 years and varies between 19 and 53 years. Most of the participants were students (80%) of diverse fields of economics. Their self-rated competence with PC (mean m = 7.66) and Internet (m = 8.18) was relatively high. Also the self-rated competence with information search via the Internet was on a high level (m = 7.33). EconDesk was known to 38.5% of the participants. EconStor was known by 31.3 % of the participants.

3.1.2 Online Subsample

The online subsample consisted of 73 persons (45 male and 28 female) who completed the survey. The drop-out rate was 21.50% (i.e., from altogether 93 persons who started, 20 persons skipped the questionnaire before completion). Their age was at average m = 29.48 years. Most of them were students (34.2%), employed (28.8%) or both (30.1%). Their self-rated competence with PC (m = 7.73), with the Internet (m = 8.08) and with information search via the Internet (m = 7.36) was rather high. EconDesk was known by 42.2% of the participants and 20.3% of the participants were familiar with EconStor.

3.2 Analyses of the Quantitative Data (SUS and SQuaLL)

The descriptive data showed that the website was evaluated rather positive, the homepage as well as the embedded three core services. However, it had to be taken into account, that all participants could be assumed as regular customers of the ZBW which in turn implies a relatively positive view of the ZBW. (People who were discouraged after the first visit of the ZBW’s homepage or EconBiz were with a high probability not in the sample, because they didn’t fulfil the prerequisite for participation.) Thus, the absolute values of the usability evaluation were probably biased in a positive way.

The comparison between the offline versus the online subsample was done by a t-test for independent groups for each of the dependent variables, i.e, the SUS-scores of the homepage and the three services (homepage-SUS, EconBiz-SUS, EconDesk-SUS, EconStor-SUS) as well as the SQuaLL (form A) as quality index for EconBiz.

No significant differences were found. None of the control variables showed any influence on the findings.
Thus, the statistical comparison provided no evidence that the assessment procedure (offline versus online) had any impact on the evaluation results. The means of the dependent variables of the offline sample versus the online sample are visualized in Fig. 1.

Overall, the quantitative results of the offline evaluation by the SUS and the SQuaLL led to an analogous evaluation like the online version. Thus, the data suggests that the offline version and the online version can be used equally for quantitative usability evaluation and benchmarking.

3.3 Analyses on the Qualitative Data

Besides the quantitative data assessment, both versions of the questionnaires comprised also the possibility to give open comments in the end. Thereby, the participants of the offline subsample used this possibility very exhaustively. Partly, they repeat their answers of the qualitative scribbling task, partly, they gave critique, laud or new hints on improvements and partly, they commented on the questionnaire and the usability study itself. Contrariwise, in the offline version nearly no participant gave any annotation. That means, heuristic insights and additional qualitative information on the homepage and the three services was received only by the offline version.

4 DISCUSSION AND OUTLOOK

In this paper, I described a comparison of an offline version versus an online version of a benchmarking questionnaire for the evaluation of the website of a Library 2.0, namely the ZBW. The assessed data referred to the ZBW's website (homepage and three core services) before the relaunch. Both studies aimed at the assessment of an evaluation baseline that can be used as comparison standard (benchmark) for future evaluations after the relaunch and further substantial changes of the ZBW's website in the future. The usability was evaluated by the short scale SUS for the homepage and the three services. Additionally, in both versions the quality of the literature search portal EconBiz was evaluated by the SQuaLL and the participants had the possibility to give additional comments in the end.
There was no evidence for a significant influence of the assessment procedure (offline versus online) on the mean values of the homepage’s evaluation and the three services with the standardized rating-scales SUS and SQuaLL. This suggests that the offline and online assessment by means of these scales can be seen as equivalent assessment procedure for evaluation and iterative benchmarking.

However, so far this result has to be limited to the used standardized rating scales (SUS, SQuaLL) and the context of website evaluation. Additionally, it has to be taken into account, that I used several arrangements to establish at least some controllability of the online sample: First, clear prerequisites for participation were stated. Second, the announcement of the online questionnaire was made on carefully selected mailing lists and webpages. Third, every participant had to insert a personally created ID which prevents (at least partly) double participations and faked answers. Fourth, participants had to provide their emails for the participation in the lottery.

It is important to note, that this comparison was a kind of field test. There are methodological limitations due to the long time interval between offline and online measurement, the different length of the two versions, and some slight changes of the website (e.g., updated texts and announcements on the homepage). However, the comparison was needed and useful in the sense that it tested how robust the evaluation results were towards this outer noise and the usual (minor) updating activities that are not directly related to the usability and quality. The line of reasoning is: If even under these circumstances (which all could theoretically lead to different results), the evaluation was on an analogous level, this is a good indicator for a robust evaluation which is independent of the practical way of assessment (offline versus online) and the normal ongoing changes of a website that are not related to the usability or the quality of services. Beyond this background, the usability indices of the SUS and the quality index of the SQuaLL can be estimated as rather robust towards the way of assessment (offline versus online), the circumstances and outer noise.

However, it is also important to note, that the online assessment delivered no additional comments or annotations of the participants. This was even more important since in the online version, the participants had only one possibility to give open feedback and critique, but they didn’t use it. Contrariwise, the participants of the offline version had several possibilities for providing feedback and one might assume that they are exhausted and tired of giving comments again at the very end. But the opposite happened: They are even more motivated to give (again) further annotations. One reason might be that the rather long offline version with the playful scribbling task had a higher stimulative nature for open (creative) comments. Another possible reason could be the face-to-face setting with an instructor. Maybe, this more personal touch enhanced the motivation of the participants to give personal comments. A third reason could be that the general motivation to participate was higher for the offline version, because the expenditure as well as the reward of participation were higher.

Overall, the findings suggest that for a pure evaluation control check a short online measurement is sufficient and leads to equal results like a long offline questionnaire. However, if deeper insights are needed, an offline assessment is more appropriate.

REFERENCES


