Teaching the evaluation of web usability

Paper 148

Abstract

This paper describes the design and use of a simple method for comparative website evaluation that has been used for the purposes of teaching web design to University students. The method can be learnt within two hours by a novice user or typical customer. The method is not dependant upon the environment being used by the tester and can be adjusted according to the subjective preferences they may have. Results are presented of the use of the method in practice in comparing the sites of a number of airlines. These suggest that the method is both sufficiently rigorous to produce reliable results, and flexible enough for users to customise. It is an effective tool in teaching the principles of web design.

Keywords

Website evaluation, user evaluation, web metrics, usability

1. Introduction

Contemporary thinking concerning website evaluation has been strongly influenced by literature in the field of software engineering. In particular this includes work concerning human computer interface (HCI) design, software metrics and software quality. Leading theorists in web usability such as Nielsen, Shniederman and Preece were previously known for their work in HCI design. Norman Fenton's work on software metrics is frequently cited in papers on web usability. The success of market led software companies has shown beyond doubt that it is the customer perspective on quality that determines success or failure. This is well illustrated in Cusumano and Selby's book 'Microsoft Secrets' (Cusumano 1999) and is enshrined in the International Organisation for Standardisation (ISO) definition of quality as fitness for purpose. Creating a fit with customer requirements is widely acknowledged as the single most important success factor in software projects. The difficulty is in establishing these requirements.

Where the interface to information systems is primarily visual, requirements can only be properly established by offering users the opportunity to test and evaluate a working version of the product. It is well established in the literature of Information Systems Design (Eason1988) that users should be involved in the development and testing of technology that they will use. Sullivan argued for a move away from a narrow conception of usability testing and towards the active participation of the user (Sullivan 1989).This idea has developed through the work of Patton (1997) in his 'utilization-focused evaluation' approach. From this background Usability Engineering has developed as a subject in its own right (Ferre et al 2001, Faulkner 2000)

2. Evaluating Web Usability

Usability is defined by ISO as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 1998, part 11). The central importance of usability to web design was firmly established in Nielsen's work 'Designing Web usability: the practice of simplicity' (Nielsen 2000). Alongside this has been the recognition that users should in turn be central to the process of the evaluation. Empirical methods for assessing usability involving real users are the most popular for web interfaces (Nielsen 1994). Travis (2003) argues that usability testing by definition uses customers in the evaluation process. He goes on to suggest that such testing requires structure given the lack of expertise of the customer. Dicks (2002) goes further stating that valid and reliable results need ‘formal methods, which essentially require large participant sample sizes, careful test construction and implementation, and analysis of inferential statistics’. Methods used for customers to test web usability essentially breakdown into two categories. On the one hand are the walkthrough methods which involve the customer testing the site in controlled conditions and recording their responses. On the other hand are the methods which involve the use of pre-defined heuristics which are deployed by the customer. Both have developed from methods used in the testing of software user interfaces (Jeffries 1991)

2.1 Walkthroughs

The walkthrough approach works best when you want a group of customers to test the design of your pre-developed site. Customer reactions can allow you to modify your design. Spool et al. (1999) conducted a study on the usability of eight well known corporate web sites. The participants of the study were given information they had to find from given sites. They were then observed on their ‘scavenger hunt’ in laboratory conditions. Their ease of finding information and reactions to this were measured by use of both observation and questionnaires. The results were in turn compared to benchmark values such as time taken to access a given page or level of difficulty in making an online transaction. Spool compared these metrics between sites in order to establish a ranking of quality. This work has since been replicated or refined many times (for example, Procter and Symonds 2001).

A variation on the walkthrough approach is a talkthrough or think-aloud (Dix 2004) method. With this method the participant is encouraged to think out loud as they use a website, whilst being observed. This process provides evidence concerning the actions and motivations of the participant. A number of means, ranging from video to computer logging, can be used to record these reactions.

Both walkthrough approaches can successfully identify weaknesses in specific elements of the site design and the achievement of key performance indicators for the site. The talkthrough can provide richer data than a simple walkthrough, but any observation of participant behaviour inevitably is very time consuming on the part of the researcher which limits the sample size and therefore generalisability of the results.

In addition, all walkthroughs have two potential drawbacks. Firstly, ‘participants feel as if though they are being evaluated, which undoubtedly affects their performance’ (Nielsen, Clemmensen and Yssing 2002) and their willingness to communicate. Secondly, the approach does not necessarily deal well with the inconsistency of user’s skill, environment, preferences or objectives. The benchmark values are determined by the researcher and not the user and thus the walkthrough can be an effective technique *in the hands of the developer.*

2.2 Heuristics testing

The term heuristics is used in this context to mean best practice usability principles. Time of loading a website homepage with a specific computer on a given connection would be an example of an heuristic test. The process of applying heuristics is very similar to the use of benchmarks in computing. The basic approach has been a) decide upon the set of heuristics you wish to measure b) apply the heuristics and c) measure the response. Most commonly, heuristics are placed in the hands of a small group of expert evaluators who independently take measurements. These measurements are then compared and contrasted. This is basically the process used when software is evaluated in the computer press. For example speed of managing a financial payment is measured for a group of websites/software packages and conclusions reached on the basis of these results. Of course, such a process is widely used in the tourism industry, for example in the allocation of star ratings to hotels.

When placed in the hands of expert evaluators in controlled conditions, heuristics can provide very accurate results for a given set of tests. Frequently the tests are very precisely calibrated. For example testers may allocate different scores according to the fractions of a second taken to access a page or according to the size/ resolution of files used on a given website. The heuristics approach can be taken further to include assessment of usability problems within specified areas. For example, an evaluator may be required to make an assessment of frequency of problems with links within a site and their severity. Travis (2003) and Dix (2004) have separately proposed similar methods that use heuristics. In each case an expert evaluator goes through the web interface screen by screen recording problems and their frequency. In Dix's method, the evaluator makes an assessment of the seriousness of the problem and combines this with a consideration of the frequency and persistence of the problem. An overall severity rating is thus achieved using a ranking of 0 (usability problem not worth acting upon) to 4 (usability catastrophe). This process is similar to the standard techniques used in project management for the assessment of risk. The heuristic approach can be adapted for website users. Users are presented with a set of criteria, eg speed of navigation, and asked them to provide a score for each site on the basis of their testing. This can be conducted in a laboratory to control the conditions of the testing.

When deployed with users the heuristic approach does have the beauty of simplicity. The researcher decides upon the heuristics to be used and their weighting, hands these over to the user and awaits the results. After this the site(s) can be modified according to user 'feedback' and the process is concluded. The method can also avoid some of the cost involved with the participation observation aspects of the walkthrough approach.

The main problem with this approach is that **what** is measured and the **value** attached to this measurement is still essentially determined by the developer. The user conducts the measurement but the developer/researcher determines what is measured. Even if the heuristics to be measured are agreed with the user when the requirements for the site are first established, the problem is that inexperienced customers are rarely clear about their requirements in advance. Truex, Baskerville and Klein, in their important paper on 'Growing Systems in Emergent Organisations' (Truex et al 1999) argue "Since the users' needs are evolving, even during the requirements determination activities, users become frustrated and trapped by the system they are helping to shape". Furthermore, whilst it is relatively easy to specify heuristics concerning technical characteristics of the site, it is difficult to do so for softer impressions of the site and the site aesthetics.

Taken to its extreme the heuristic approach can lead to the dialectical position whereby the user is asked to test those elements of design that have been foremost in the developers mind and thus return a positive result for a poorly designed site.

3. Creating a usable evaluation method

3.1 Software package evaluation

This dilemma of trying to establish a method that gives the user control, whilst recognising their limited expertise, has been experienced in software design over many years. The problem became easier to solve when the emphasis on software purchasing switched from bespoke systems to off the shelf packages. Users had the opportunity to view the final product prior to purchase and compare with similar products. All they now needed was a framework in which to make this comparison.

Sharland (1991) developed a simple framework for software users to evaluate and compare software packages with a view to purchase. The framework firstly required users to consider their purchasing criteria, for example cost, speed, fit with existing systems. Possible criteria were presented to the user to take account of their lack of expertise. This allowed them to consider important issues such as product support and company reputation which may not have been considered at first. Secondly the user was required to decide the relative importance of these criteria. Finally they viewed the packages in order to score the criteria according to a fixed scale.

This approach to some extent addresses the limitations of the walkthrough and heuristics methods outlined above and has been adapted and refined for the purpose of evaluating web usability from the user perspective. The basic approach is a modification of the heuristics approach and can be summarised as follows:
a) in discussion with the user decided upon the design criteria you wish to evaluate

b) weight these according to the context and according to user preferences
c) apply these to the given sites
d) compare the scores between the sites
e) use the exercise for the purpose of comparing similar website and in order to develop understanding of requirements

3.2 Establishing web design criteria

A substantial body of literature has developed which describes key factors in web usability and web site design. Possibly the best known is the work of Jakob Nielsen (2000). Various authors have reported their work in creating metrics for usability and site design such as Becker & Mottay (2001), Spool et al (1999), Palmer (2001) Furthermore, numerous commercial organisations such as Reinken.com, Surveysite.com, Gomex.com, Bizrate.com and WebDesign.about.com have developed their own frameworks for the same purpose. A synthesis of these sources established seven criteria which cover the main areas of web site design. It must be emphasised that this list can easily be adjusted and is not central to the idea of the method.

The criteria are listed below with sub lists of questions that indicate to the user the meaning of each:

1. Superstructure

- Is the site layout easy to understand
- Is the navigation from page to page easy
- Is there an intuitive feel for the visitor
- Is it easy to manoeuvre back to the home page or top of the page
- Is the loading time quick and efficient

2. Graphics

- Are they clear and attractive
- Are they necessary to the page
- Are they distracting or excessive
- Do they contribute to understanding
- Will they contribute to excessive loading time
- Do they aid the visitor with site navigation

3. Use of colour

- Are the colours attractive to most visitors
- Are there too many colours which look messy
- Would more colour enable the visitor to understand the content more
- Do the colours follow web standards and user expectations (e.g. link colours)
- Has colour blindness been considered

4. Content

- Is the content interesting and of value to the user
- How good is the interaction/ how ‘rich’ is the content
- Is it regularly updated
- Does it attract visitors
- Is it necessary and in good taste
- Is it fun/ good style/ personality

5. Readability

- Are the pages easy for the visitor to read
- Are the fonts readable, attractive and properly sized
- Are the pages in a logical sequence
- Will the site look attractive and fit with different browsers/ screens

- Do the graphics add to/ detract from readability

6. Page Layout

(Concerned with the structure of each page and not the site as a whole)

- Is the design for each page consistent
- Does it conform to use expectations/does the user have much control
- Can we assess the template that has been used by the designer
- Is there sufficient space for content – do you like it

7. Hyperlinks

- Do the links enhance the purpose of the site/ are they explained/ do they have the ‘click here’ problem
- Do the links lead the visitor away from the site
- Are the links easy to spot/ standard colours/ consistent/ is there a strategy
- Are internal links used to avoid excessive scrolling
- Can the user get lost/ can we get back to the home page

3.3 Deciding the relative importance of the design criteria

The user is presented with a table containing the criteria as shown in figure 1:

Criteria	**Weight**
Superstructure	
Graphics	
Colour	
Content	
Readability	
Page Layout	
Links	

Fig. 1

For each one a value in the range 1 to 5 must be assigned. A weight of 1 indicates negligible importance and a weight of 5 indicates a vital attribute or criteria. The weight remains a constant for the whole duration of an evaluation/ comparison of similar web sites.

3.4 Assigning values and calculating score

Criteria	**Weight**	**Value for A**
Superstructure		
Graphics		
Colour		
Content		
Readability		
Page Layout		
Links		

Fig. 2

The user then views the first website (site A) and assigns a value for each criteria using the scale 0 to 5. A value of 0 would indicate very poor and a value of 5 excellent.

Criteria	**Weight**	**Value for A**	**Score for A**
Superstructure	4	5	20
Graphics	2	3	6
Colour	3	3	9
Content	4	0	0
Readability	5	1	5
Page Layout	1	4	4
Links	2	5	10
			Total 54

Fig. 3

Thirdly a score for each criteria for site A is calculated by multiplying the weight with the value as indicated in the example above. Finally, the scores are added together to establish a total score for the site. In the example above the total for site A is 54.

Next the user views site B. Values are then assigned for site B and scores calculated. This can be repeated for as many sites as need to be compared.

Finally, the different site scores can be compared to get an indication of the best design. The number of sites tested is at the discretion of the evaluation designer.

4. Implementation

The method was tested on five occasions with groups of undergraduate students using desktop PCs who were given 6 airline websites to evaluate and compare. Prior to this the method was tested with two groups who used the method to evaluate and compare the sites of a group of hotel chains.

The criteria were explained to the students with the help of examples. Where necessary, guidance was given. For both expert and non-expert users alike, one hour proved adequate for this explanation. They then sorted themselves into pairs for the purpose of conducting the exercise. Once they were clear about the criteria they were to measure, the method of measurement was explained. It was stressed that allocation of weightings was relative. For example, if colour was considered more important than content it should be given a greater weighting. The pair had to come to an agreement on this. They learnt that evaluation criteria are context specific. For example colour may be weighted more heavily when evaluating museum web sites than when evaluating those of airlines.

Once each pair had agreed upon weightings they were given specific sites to measure values for, using the table above. For a set of 6 web sites around 1 hour was required for this stage. At the end of this time the users could combine the weight and value for each of the criteria in order to achieve a score for each of the sites. Simple addition gave a total score for each site enabling users to rank the sites. Following completion of the exercise the results for each pair of students was tabulated.

5. Results

The results for the five tests of the evaluation framework are shown below as group 1,2,3,4 and 5. On each occasion the same six airline websites were compared. There were no problems with access to any of the specified sites on any occasion and the students were advised that they should not consider issues such as price or availability of flights in their judgement of usability. On each occasion, with students of different abilities and group sizes the entire exercise took no more than two and half hours from initial explanation to testing to joint discussion of results. 342 students in total took part ie 171 pairs. A simple table is presented below showing the sites gaining the highest/lowest score from the five groups:

Note that in some cases two or more sites were ranked as being equivalent in which case a score has been entered in each cell.

Student group 1 n=18

Site	No. with highest score	No. with lowest score
Site A	1	6
Site B	2	1
Site C	4	3
Site D	8	0
Site E	4	2
Site F	0	8

Student group 2 n=47

Site	No. with highest score	No. with lowest score
Site A	3	7
Site B	4	10
Site C	5	5
Site D	21	3
Site E	13	5
Site F	3	17

Student group 3 n=20

Site	No. with highest score	No. with lowest score
Site A	2	4
Site B	2	6
Site C	4	0
Site D	8	1
Site E	3	3
Site F	1	6

Student group 4 n=36

Site	No. with highest score	No. with lowest score
Site A	1	10
Site B	1	8
Site C	4	1
Site D	17	3
Site E	9	3
Site F	4	13

Student group 5 n=50

Site	No. with highest score	No. with lowest score
Site A	5	13
Site B	5	8
Site C	6	4
Site D	24	3
Site E	11	3
Site F	3	19

Student groups 1-5 Combined n=171

Site	No. with highest score	No. with lowest score
Site A	12	40
Site B	14	33
Site C	23	13
Site D	78	10
Site E	40	16
Site F	11	63

There was significant consistency in the results, with every group ranking site D as having the highest score; a total of 78 pairs out of 171. In every group site F had the lowest score; a total of 63 pairs our of 171. In group 3 this was ranked lowest jointly with site B. There was also substantial divergence. For example, although it was widely agreed that site D was the most usable, ten pairs of students ranked it as the least usable. It is recognised that individual results are not necessarily independent since student pairs had the opportunity to confer whilst conducting the exercise.

Prior to conducting the exercise the students were asked to consider whether they felt that web evaluation was primarily a subjective or objective issue. The overwhelming majority felt that it was subjective and that the quality of a site was a matter of taste. Following the collation of results the students were asked to return to this issue. Although no detailed data has been recorded for this question, the majority of pairs wished to modify their original view, given the fact there was significant agreement on the relative quality of the design of the sites which had been evaluated. At the same time conducting such an exercise led to their realisation of the limits of objectivity of any evaluation framework. For example, one student regarded green as the ideal background colour for a web site, whilst others were disdainful. One regarded a choice of languages on the home page as really useful whilst another regarded it as an unnecessary distraction. In conclusion most students wished to acknowledge that evaluation is a combination of subjective and objective opinion. The results demonstrated to them that although there were differences of opinion, there was also significant agreement on aspects of good design practice (for example fast download time). This they felt had resulted in approximately half the students agreeing on the best designed/ worst designed sites.

One student commented "The exercise gave students some much needed structure to their evaluations, providing pointers to look out for". Another that “it allowed us to work out which we felt were the most important things to look for”. One of the students commented "How good is good? This is an issue for inexperienced web site reviewers as they struggle to quantify standards and make comparisons in web design. All in all (this is) an excellent way to get students to start thinking in the right manner about the salient points of web design".

6. Discussion

Those taking part in the exercise on each occasion entered a lively debate concerning the criteria that should be measured and the weightings that should be attached. Thus one commented "I learnt that there are different views on what are key criteria. It also helped me to form my own views on what should be evaluated".

Three additional criteria have been proposed on more than one occasion. These are:

a) Accessibility of the site. Accessibility is concerned with the usability for disabled users and especially those who are visually impaired. It is also concerned with conformance to Web content accessibility

b) Quality of interaction on the site as a criterion in its own right. Quality of interaction is concerned with the simplicity of making a booking/payment and information provided. It may also take into account the perceived safety of the process.

c) Promotion of the site. Guidelines and coding standards issued by the World Wide Web consortium. Promotion is concerned with the ease of finding the site, the optimisation of the site for retrieval by search engines and a less easy notion to explain, the site ‘stickability’.

Incorporating these into the evaluation framework in practice is difficult since they require either specialist users, a specialist environment or specialist skills.

Some students also felt that the range of values available for measurement purposes, i.e. 0 to 5, was insufficient, and have subsequently reused the framework on different sites using a range of 0 to 10: the level of gradation depends upon the expertise of the evaluator. Indeed, the use of an empirical method for assessing personal views of usability is in itself a crude technique.

Further work would be needed in different circumstances, with a greater mix of users using different equipment in order to be able to generalise the results. For example, the exercise reported may have led to completely different results if conducted with mobile devices. Whilst the numbers presented do suggest that the framework can produce meaningful results in controlled conditions, they are not the most important aspect of the method. The most important achievement of the use of the framework has been in developing the understanding of those who have used it: in this case 5 different groups of students. Use of the framework can enable novice users to develop their understanding of usability and thus of their own requirements.

7. Conclusions

The paper has presented a method of conducting web evaluation which is user centred. It is founded on an analysis of the literature of user centred web design combined with a system of measurement similar that that which has been used for many years in other situations such as risk analysis and software package evaluation. This builds upoon established methods of user centred web evaluation, namely walkthroughs and heuristics. Results have been presented of the use of the method in practice for the purposes of teaching the evaluation of web usability to undergraduate students. The method was tested with 5 groups of students and there was sufficient agreement amongst the different groups to suggest that it is a worthwhile framework for web evaluation that can be used for comparing similar web sites. Further research conducted in different circumstances and with different users would be needed to make any claim of scientific proof. More importantly we contend that use of the method can develop the understanding of web usability on the part of the user and thus empower them to make better informed choices when establishing their requirements.

Effective evaluation requires some basic expertise. The instrument described is a simple framework for users to both evaluate and compare a group of similar web sites. The process of evaluation is an effective way of learning about both web design and evaluation itself.

References

The author gratefully acknowledges the use of the MSc dissertation ‘Researching ECommerce Usability’ (2004 unpublished) by Jiten Makan in the literature review section of this paper.

Becker S. & Mottay F. *A global perspective on web site usability* IEEE Software Jan/Feb 2001
Cusumano M.Selby Mi*crosoft Secrets*: *How the World's Most Powerful Software Company Creates Technology, Shapes Markets, and Manages People* Profile 1997
Dicks S.R. (2002) *Mis-Usability: on the uses and misuses of usability testing* Proceedings of the 20th Annual International Conference on Computer Documentation Ontario, Canada Oct 2002
Dix A. (et al) (2004) *Human-Computer Interaction* 3e Pearson Education Limited/Prentice Hall, Essex
Eason K. (1988) *Information Technology and Organisational Change:* Taylor and Francis
Faulkner X (2000) *Usability Engineering*, Basingstoke: Palgrave
Fenton N. (1994) *Software measurement – a necessary scientific basis* IEEE Trans. Software Engineering 20(3) 199-206
Ferre X. , Juristo N. , Windl H. , Constantine L. *Usability Basics for Software Developers* IEEE Software Jan/Feb 2001
ISO (1998) *Ergonomic requirements for office work with visual display terminals* ISO 92411-11, ISO, Geneva
Jeffries R (et al) (1991) *User Interface evaluation in the real world: A comparison of four techniques* Proceedings of SIGCHI Conference on Human Factors in computing: Reaching through technology. Louisiana USA March 1991
Nielsen (1994) *Usability inspection methods* Conference on Human Factors in Computing Systems, Boston USA
Nielsen, J. (2000) *Designing Web Usability,* Indianapolis: New Riders Publishing
Nielsen, Clemmensen and Yssing (2002) *Getting access to what goes on in people's heads? – Reflections on the think-aloud technique* Proceedings of the 2nd Nordic Conference on Human Computer Interaction Aarhus, Denmark Oct 2002
Palmer (2002) *Web site Usability, design and performance metrics* Information Systems Research 13(2) pp.151-167
Patton M. (1997) *Utilization-focused evaluation,* London: Sage
Procter C. and Symonds J. (2001) *Designing for Web site usability* The Australian Journal of Information Systems, vol 9, no 1
Salmon G. (2000) *E-moderating: the key to teaching and learning online,* London : Kogan Page, 2000
Sharland R. (1991) *Package Evaluation: a practical guide to selecting applications and systems* Aldershot: Avebury
Spool, J.M., Scanlon, T., Schroeder, W., Snyder R. and DeAngelo T. (1999) *Web Site Usability: A Designers Guide*, San Francisco: Morgan Kaufmann Publishers.
Sullivan P. (1989) *Beyond a narrow conception of usability testing* IEEE Transactions on Professional Communication 32(4) December
Travis D. (2003) *E-Commerce usability: Tools and techniques to perfect the on-line experience,* Taylor and Francis Group Publishers, London UK
Truex, D. P., R. Baskerville, and H. Klein: 1999, *Growing systems in emergent organizations*. Communications of the ACM 42(8), 117–123.