What Do Users Really Care About? A Comparison of Usability Problems Found by Users and Experts on Highly Interactive Websites

Helen Petrie
Human Computer Interaction Research Group
Department of Computer Science
University of York, York UK YO10 3WF
Helen.Petrie@york.ac.uk

Christopher Power
Human Computer Interaction Research Group
Department of Computer Science
University of York, York UK YO10 3WF
Christopher.Power@york.ac.uk

ABSTRACT
Expert evaluation methods, such as heuristic evaluation, are still popular in spite of numerous criticisms of their effectiveness. This paper investigates the usability problems found in the evaluation of six highly interactive websites by 30 users in a task-based evaluation and 14 experts using three different expert evaluation methods. A grounded theory approach was taken to categorize 935 usability problems from the evaluation. Four major categories emerged: Physical presentation, Content, Information Architecture and Interactivity. Each major category had between 5 and 16 sub-categories. The categories and sub-categories were then analysed for whether they were found by users only, experts only or both users and experts. This allowed us to develop an evidence-based set of 21 heuristics to assist in the development and evaluation of interactive websites.

Author Keywords
Expert evaluation; heuristic evaluation; user evaluation; usability problems; heuristics.

ACM Classification Keywords
H.5.2. [User Interfaces]: Evaluation/methodology

General Terms
Experimentation, Human Factors.

INTRODUCTION
Expert evaluation is a logical and important component in the development of interactive systems. It makes sense to have experts identify problems with a system before exposing users to it, even if those users are only conducting an evaluation themselves. In the case of rapidly developed and deployed systems, such as many websites, expert evaluation may be the only evaluation that is undertaken before a system goes live. There are numerous forms of expert evaluation, including Cognitive Walkthrough [15, 16], Guidelines Review [6] and Consistency Inspection [14], but the best known is Heuristic Evaluation (HE), developed by Molich and Nielsen [17, 18, 20]. HE involves asking 3 – 5 usability experts to work through an interactive system, seeing whether any of a set of heuristics is violated, thus creating usability problems. The experts then come together and produce a consolidated set of usability problems and rate them on a 4 level severity scale from "catastrophic" to "cosmetic only". HE is described in numerous HCI textbooks [6, 14, 22] and on authoritative websites such as usability.gov and UsabilityNet. Many of these sources quote the original Molich and Nielsen heuristics, as well as Shneiderman's 8 golden rules of interface design [23] and Tognazzini's basic principles for interface design [25] which can also guide an HE.

HE and other forms of expert evaluation have come in for a range of criticisms, including:

- low overlap between usability problems proposed by expert evaluations and user evaluation, as low as 10% [1, 9, 12]
- different experts or groups of experts produce different problem sets [8, 9, 12, 13]
- expert evaluations over-emphasize low severity problems at the expense of high severity problems [12]

However, the studies that have compared different evaluation methods have also come in for considerable criticism, particularly by Gray and Salzman [7] and particular aspects of the comparisons, such as matching problems between methods have also been criticized [5, 10].

In this paper, we take a different approach to the issues concerning expert evaluation. The criticisms of expert evaluation methods have not deterred researchers and practitioners from using these methods. But if the overlap between problems found by experts and those reported by users is only of the order of 10% and these methods may not identify the problems that users find most severe, then we must ask if expert evaluations are a good use of the time and effort of the experts and the development teams? If we

CHI'12, May 5–10, 2012, Austin, Texas, USA.

want to keep using these methods, what can be done to improve the effectiveness of these methods? In this paper we will concentrate on this question, and examine sets of heuristics used by experts in evaluations.

The first concern is with the continued use of the original Molich and Nielsen [17, 18, 20] heuristics to guide evaluation and their unvalidated adaptation for specific areas of modern interactive systems, particularly the web. The Molich and Nielsen heuristics were based on sound evidence and were a sensible attempt to cut through the complexity of interface guidelines available at the time of their writing. Molich and Nielsen [17] ran a competition on the evaluation of an interactive system and analysed 77 entries for the types of usability problems produced in them. Nielsen [19] analysed 249 usability problems from 11 different projects to validate the heuristics. Therefore it is clear that Molich and Nielsen were careful in developing the heuristics from a sound evidence base.

However, since then, interactive systems have become much more complex and diverse. In addition, the web has a particular set of conventions and methods of interaction to which users have become accustomed during regular use. These conventions and methods have become ingrained in users' mental models about how the web works, and thus will influence what they perceive as a usability problem. The web also has content and information architecture that may result in usability problems that were not typical of 1980s interfaces. Thus, the Molich and Nielsen heuristics may no longer capture the main usability problems that user have.

In response to the evolution of interactive systems to web-based systems, a number of authors have adapted the Molich and Nielsen heuristics for the web. Instone [11] produced "site usability heuristics for the web" and more recently Budd [3] produced "heuristics for modern web application development". Each of these is a light reworking of the Molich and Nielsen heuristics with examples drawn from the web. However, the question remains: are these heuristics really representative of the problems that users have with current interactive systems, in particular with web-based systems? It does not seem appropriate to take the problems that users had with interactive systems in the late 1980s and transfer them to web-based systems in the 2010s. This is particularly problematic when there is no empirical evidence that these problems are actually the ones encountered by users. Is this in fact one of the reasons that expert evaluations are producing so little overlap with user evaluations – that the heuristics used by experts are not fit for purpose?

The paper will explore this question by comparing the usability problems found in a large-scale evaluation of six complex, highly interactive, government websites. The websites were evaluated both by potential users and by experts using three different expert evaluation methods. However, it is not the intention of this paper to compare between the problems found between the three expert methods, given the problems of comparing evaluation methods [7]. Instead, in this paper we are investigating what types of usability problems users encounter that are missed by experts and vice versa. From this analysis, we will be able to propose a current, evidence-based set of heuristics to guide developers and expert evaluators of highly interactive websites.

METHOD

Design

Six complex, highly interactive websites were each evaluated by 15 potential users using a think-aloud protocol and three different expert evaluation methods using teams of 3 experts. The three expert evaluation methods used were Collaborative Heuristic Evaluation (CHE) [21], Group Expert Walkthrough (GEW) and Group Domain Expert Walkthrough (DEW) [14].

Both potential users and experts were asked to identify usability problems and rate them on a four-point scale from "catastrophic" to "cosmetic" [18]. A grounded theory approach [2, 24] using open coding was used to categorize the usability problems found in the different evaluations, to allow natural groupings of problems to emerge. Problems found by experts only, users only and both users and experts were then analyzed.

Websites

The websites used were highly interactive and transactional government websites. Several of the websites were ones where users provide information to government agencies, for example to qualify for particular benefits. Several others were informational, but required the user to find information by specifying criteria for searching and for filtering results. A number of the websites were already publicly available at the time of evaluation, while others were fully functioning prototypes, ready to launch. For reasons of confidentiality, the websites will not be named and will be referred to as website A – F.

Method for user evaluations

Participants

30 participants took part in the study, 13 were women and 17 men. Participant ages ranged from 22 to 61 years, with a mean age of 33.3 years (standard deviation = 10.61). On average participants had used the web for 11 - 15 years and rated their web experience as "High" on a five-point scale. Participants reported average daily web use of between one and five hours per day. None of the participants had previously used any of the websites evaluated in the study.

16 participants were university students and 14 worked in a range of occupations. Three of the websites related to higher education and initial career plans, so these websites were evaluated by the students, as they would be target users for those websites. The other three websites were aimed at the general population, so the other participants evaluated these.

26 participants evaluated three websites each and four participants evaluated two websites each. Participants were remunerated with Amazon gift vouchers, £5 per website evaluated.

Equipment

Standard personal computers running either Windows or MacOS and a range of web browsers (Internet Explorer, Firefox) were used, according to the individual participant's preferences. The computer also ran a screen capture program (Morae for the Windows machines, ScreenFlow for the MacOS machines) that recorded the screen and voice of participant and researcher.

Procedure

Each session lasted 60 or 90 minutes, depending on the number of websites the participant chose to evaluate. Participants were first briefed about the study and signed an informed consent form. They then completed a brief demographic user questionnaire. For each website evaluation, participants were provided with a persona and a scenario of use and the relevant information needed to complete the scenario. Participants undertook a concurrent think aloud protocol, talking though the usability problems as they were encountered. Participants were gently prompted if they did not keep up the think aloud commentary.

Each time a problem was encountered i.e. if the participant made some comment that indicated a problem (e.g. "I don't understand this" "I can't figure out what to do now"), the researcher asked the participant to pause briefly and rate the problem for its severity on a scale where 1 = cosmetic, 2 = minor, 3 = major, 4 = catastrophic.

Participants were allowed 30 minutes to evaluate each website. This procedure was repeated for two or three websites. Participants then completed a brief post-study questionnaire, were debriefed and signed off the informed consent form.

Method for expert evaluations

Each group of usability experts evaluated all six websites, with each expert evaluating each website only once. Each group used each method twice and each website was evaluated with each method three times. This design meant that individual differences between experts and between the websites would not have undue effects on the results.

Experts

14 usability experts participated, five were women, nine were men. The majority had higher education qualifications or had taken courses in HCI. The majority had over five years experience in usability and worked as professionals in user experience, interaction or software/product design with usability making up between half and all of their current role. 11 described themselves as "experienced" in usability, the remainder as "junior". Nearly all had conducted HEs but only three had previously participated in GEWs and DEWs. 11 of the usability experts worked for DirectGov (the UK government's digital service, providing online access to a wide range of government services and information) or organizations that provided usability services to DirectGov. Three of the usability experts worked for the University of York.

Six domain experts participated: four were women, two were men. Two were business analysts who occasionally provided help in relation to user issues, others were team leaders or advisors responsible for re-design of digital services and regularly provided information about users to development teams. Half the domain experts had less than one year's experience in the particular domain, the others had two to four years experience. Only one domain expert had participated in a usability evaluation before this study.

Equipment

Two computers were used in each expert evaluation session: one accessed the website being evaluated and was under control of the experts; the other was used to record problems raised by the experts. The displays of both computers were projected onto a wall so all the group could see it clearly.

Overall Procedure

Each evaluation session was led by a facilitator and assisted by a scribe (one of the authors in each case, neither of whom participated in the evaluation). The facilitator introduced the method and briefed experts on the procedures to be followed, including the use ofa persona, a scenario of use and provided copies of the original Molich and Nielsen heuristics and severity rating scale. Once the introduction was complete, the evaluation started and continued for exactly one hour.

Procedure for Collaborative Heuristic Evaluation (CHE)

Experts worked as a group, with one expert "driving" the website. The collaborative version of heuristic evaluation developed by Petrie and Buykx [21] was used. Any expert could propose a potential usability problem. Experts described each usability problem so the scribe could record them and to create consensus on the description of the problem. Experts then rated its severity privately using the four point rating scale [18]. If an expert did not think the potential problem was a usability problem, they rated it as having a severity of zero. This allowed different experts to provide both their view of whether a potential problem was actually a usability problem and what its severity was.

Procedure for Group Usability Expert Walkthrough (GEW)

The evaluation period was split into two 30 minute periods. For the first period, one expert took on the role of the user described in the persona and worked through the scenario of use while providing a concurrent verbal protocol. The other experts could ask questions of the "user" and note usability problems. In the second period, experts worked as a group to identify usability problems by reworking the scenario of use. Severity ratings for each usability problem were reached by consensus.

Category	Examples
PHYSICAL PRESENTATION	
Page does not render properly	Navigation is longer than the footer and overlaps the footer (A046) Text on buttons does not transform gracefully when resized (E106)
Poor, inappropriate color contrast	Dates hard to read - grey on grey background, not very clear format (A151) Interface colors are too dark, too saturated, "glowing" (C23)
Text/ interactive elements not large/clear /distinct enough	Radio button area / sensitivity is too small (D17) Date boxes too small (B50)
Page layout unclear/confusing	The heavy red bar below number of courses separated the page in two, thought the information below was something irrelevant like a advert (C83) Text very tight up to border, difficult to read, unattractive (E4)
Timing problems	Holding error message was too brief to read (A76) Clicking on 'XXX' button took more than one minute to load (D15)
Key content/ interactive elements, changes to these not noticed	Did not notice the second set of input requirements relating to the password (D5) Did not see list expansion (F5)
"Look and feel" not consistent	Page looks different - buttons have moved, text is smaller (A282) Table unexpectedly transposed - University was row, now column (F103)
CONTENT	
Too much content	Overly wordy, but nothing to assist the user (A037) Number of results is too large for user to work with effectively ("Off putting", "Overwhelming") (F64)
Content not clear enough	More plain English needed on Welcome Page (E65) Disclaimer not clear - which relevant organizations? (B48)
Content not detailed enough	The information provided was very sparse (D2) Lack of information about syllabus/ curriculum (F32)
Content inappropriate or not relevant	Insensitive explanation of terminal illness (A144) Information on overview page does not seem relevant to task (F50)
Terms not defined	Acronyms f/t and f/d not clear, leaves user to determine meaning (F48) "Includes flexible start dates" - what does this mean? (C56)
Duplicated or contradictory content	Inconsistent information about sending personal items (A262) Agreement statistics and number of respondents columns don't agree (F104)
INFORMATION ARCHITECTURE	
Content not in appropriate order	Why is all this information presented before the registration? (E76) Check list at wrong end of the process (B116)
Not enough structure to the content	A lot of information - expecting step by step process (A78) The results get lost in all the other text (D70)
Structure not clear enough	Table in pseudo-alphabetic order, unclear (F111) After completing the questionnaire, expected to go straight to the reports, not back to the beginning (D86)
Headings/titles unclear/confusing	Page is not actually university details – misleading title (F95) Help – but this is not help, this is further information (B161)
Purpose of the structures not clear	What are these boxes on the side for? (A64) Are the colors significant (block colors behind the groups of services) (A255)
INTERACTIVITY	
Lack of information on how to proceed and why things are happening	No guidance on how to use the university search (F121) Confusing that it is recovering the death certificate rather than me entering information from the death certificate (B27)
Labels/instructions/icons on interactive elements not clear	Unclear what "special notes" checkbox will do (A187) Asterisks here appear to mean incomplete, not the usual mandatory (B42)

Duplication/excessive effort required by user	Have to provide details again, even though already provided them (A190) "Please access urls shown before submitting info" - expected to do quite a lot of work (copy, paste, look up info) (B149)
Input and input formats unclear	Postcode input not at all obvious (D55) I would prefer to enter months via names than numbers (B10)
Lack of feedback on user actions and system progress	Not clear whether the system has saved results (D67) The search returned no results, provided no guidance as to why this might be (C3)
Sequence of interaction illogical	Why is the exit button at top? (B25) Activation code retrieval out of sequence (E82)
Options not logical/complete	What if I do not have UK qualifications – no options (D117) Why isn't niece/nephew in the list of relatives? (B41)
Too many options	Far too many options, when the main goal is to view a report (D83) Bank account options - too many options (A165)
Interaction not as expected	Breadcrumb trail literal, not the structure of the site (F108) Tabbing is illogical (skips) (B3)
Interactive functionality expected is missing	No way to sort short list (F65) Surprised that there was not a de-select option, considering the amount of checkboxes that were pre-selected (C2)
Links lead to external sites/are PDFs without warning	Unclear if links on this page will lead to external sites (F12) Was given a PDF doc, did not indicate this and gave no other options (B123)
Interactive and non-interactive elements not clearly identified	Why isn't Contact Us a link in the text? (E6) Arrow in table header not clearly indicated as selectable for sort functionality (F55)
Interactive elements not grouped clearly/logically	Next button is a long way away from the text I am to read (A74) Radio button for correction of error way at bottom (B98)
Security issues not highlighted	No information about how personal data is treated (A263) User unclear who will get and use the information (B36)
Problems with choosing and validating passwords	Why is password choice so restrictive? (A122) Password case-sensitive with no indication this is the case (F72)
Error messages unhelpful	Error message does not indicate what bit of information is wrong (B19) Unhelpful error message "Form percentage must be equal to "100" (E34)

Table 1. Categorization of usability problems (Axx – Fxx refer to the six websites and problems codes).

Procedure for Group Domain Expert Walkthrough (DEW)

Initially, domain experts were given a 30 - 45 minute introduction to the principles of usability evaluation. The method was then the same as that described for GEW.

The two groups of domain experts each evaluated one website in the domain of their expertise using the DEW method.

Data analysis

For each website, a unified list of usability problems from all the methods was created. A strict procedure was followed for matching problems from different methods. The problem needed to be about the same interactive element/unit of content and describe the same type of problem for the user.

A grounded theory approach was then taken to categorizing the usability problems. This was done blind to which method had produced the problems. Two researchers used an open coding technique, repeatedly summarizing and grouping the problems, until natural and appropriate categories emerged. The grounded theory also resulted in grouping the initial set of categories into more abstract categories (henceforth referred to as "major categories") such as "Physical presentation" and "Content" (see Table 1). Inter-coder reliability was then established by having a third researcher categorize a sample of 50 problems. Cohen's Kappa (K) [4] was calculated on agreement between one of the original coders and the new coder for both major category and sub-category. Both calculations showed satisfactory levels of agreement (for major categories: K = 0.93; for sub-categories: K = 0.89).

RESULTS

A total of 947 distinct problems were identified. 12 were discarded as not being usability problems (e.g. user forgot their postcode) or being too vague for categorization. This

left a pool of 935 usability problems, an average of 155.8 problems per website (standard deviation = 66.1, range 81 to 271).

Table 1 shows the emergent categorization of usability problems, for categories with five occurrences or more. This accounted for 907 problems. The four major categories that emerged from the coding were: Physical Presentation, Content, Information Architecture and Interactivity. We would have preferred to use the term "Content Architecture" to be consistent with our use of the term "Content", but as Information Architecture is a known term in HCI and beyond, that was used. Interactivity was the largest category, with 16 sub-categories; the other categories have a more even breakdown, with Physical Presentation having seven sub-categories, Content having six sub-categories and Information Architecture having five sub-categories.

Table 2 shows the distribution of problems into the major categories, for problems found by users only, experts only and both users and experts. The distribution of problems found only by users and only by experts was very similar, with no significant difference between them (chi-square = 7.45, df = 6, n.s.). Nor was there a significant difference between these distributions and the distribution of problems found by both users and experts (chi-square = 5.439, df = 3, n.s.).

Category	Users only	Experts only	Both users and experts	Total
Physical Presentation	13.4% (67)	11.2% (31)	8.5% (11)	21.0% (109)
Content	17.0% (85)	22.7% (63)	21.7% (28)	19.4% (176)
Information Architecture	8.6% (43)	10.5% (29)	8.5% (11)	9.2% (83)
Interactivity	61.1% (306)	55.6% (154)	61.2% (79)	59.4% (539)
Total	501	277	129	907

Table 2. Usability problems identified by users only, experts only and both users and experts (% and number).

However, the distribution of problems into the sub-categories was significantly different for problems found by users only and problems found by experts only for three of the four major categories: Physical Presentation (chi-square = 14.18, df = 6, $p < 0.05$), Content (chi-square = 11.78, df = 5, $p < 0.05$), Information Architecture (chi-square = 1.38, df = 4, n.s.), Interactivity (chi-square = 41.50, df = 16, $p < 0.001$).

Sub-Category	Ratio users : experts
Expected Ratio	**1.81 : 1**
Timing problems	9 : 1
Security issues not highlighted	8 : 1
Page layout unclear/confusing	5 : 1
Interactive functionality expected is missing	4.8 : 1
Input and input formats unclear	4 : 1
Links lead to external sites/are PDFs without warning	4 : 1
Poor color contrast	4 : 1
Interaction not as expected	3.7 : 1

Table 3. Usability problems that were reported more frequently by users only than experts only.

Sub-Category	Ratio experts : users
Expected Ratio	**0.55 : 1**
"Look and feel" not consistent	2 : 1
Content not clear enough	1.4 : 1
Key content/interactive element, changes to these not noticed	1.2 : 1
Headings/titles unclear/confusing	1 : 1
Purpose of the structures not clear	1 : 1
Terms not defined	1 : 1

Table 4. Usability problems that were reported more frequently by experts only than users only.

To explore where the differences in the distribution of problems in the sub-categories lay, the sub-categories for which users only found problems and those for which experts only found problems were analysed. As users found nearly twice as many problems as experts (501 vs 277, a ratio of 1.81 : 1), the ratio of user only problems to expert only problems was calculated for each sub-category. Table 3 shows those sub-categories with the most extreme ratios in favor of users only. Thus users are far more concerned with timing problem, security issues and confusing page layout. Table 4 shows those sub-categories with the most extreme ratios in favor of experts only (to make this easier to understand, the ratio of experts only : users only problems was used in this table, in comparison to an overall ration of 0.55 : 1). Thus experts are more concerned with

consistency of the "look and feel" of a website, with content not being clear enough for users to understand, and the saliency of key content and interactive elements on the page and changes to these content and interactive elements. Table 5 shows the sub-categories with the highest proportion of problems found by both users and experts. Thus both users and experts are concerned about having too many options in interaction, problems with choosing and validating passwords and interactive elements and their associated labels and text not being grouped together clearly and logically.

Sub-Category	% of problems in sub-category
Too many options	40.0% (2/5)
Problems with choosing and validating passwords	30.0 (6/20)
Interactive elements not grouped clearly/logically	27.3 (6/22)
Interactive and non-interactive element not clearly identified	26.3 (5/19)
Labels/instructions/icons on interactive elements not clear	24.0 (18/75)
Lack of feedback on user actions and system progress	24.0 (3/14)
Content not detailed enough	22.8 (13/57)
Content inappropriate or not relevant	21.2 (7/33)

Table 5. Usability problems that were reported by both experts and users.

Finally on the basis of the analysis of usability problems, we can propose a new set of heuristics for developing and evaluating current highly interactive websites. For these heuristics we looked both at the severity of problems for users and their frequency. Severe problems clearly need to be identified early if possible; but frequent problems, even if not severe, should be addressed, as the cumulative effect of many problems may also be highly detrimental to users. Therefore, usability problem sub-categories with median severity ratings from users of 2.0 or higher were identified as well as sub-categories with a problem frequency of 10 or more instances from users or both users and experts. These sub-categories are shown in Table 6, now turned into positive heuristics for developers. 18 sub-categories were identified by each method, but only 12 sub-categories were identified by both methods. Five sub-categories were identified on the basis of the mean severity of problems in that sub-category, and six sub-categories were identified on the basis of the frequency of problem in that sub-category. This information is also summarized in Table 6.

DISCUSSION AND CONCLUSIONS

This paper has presented an analysis of 935 usability problems found in the evaluation of six complex, highly interactive websites, using both user evaluation and three different expert evaluation methods. As has been found previously [1, 9, 12], the overlap between problems found by users and by experts was relatively low, in this case only 14.2%. However, in this paper the aim was not to compare usability evaluation methods, but to look at the types of problems encountered by users but missed by experts and vice versa, in order to propose a new evidence-based set of heuristics to guide both developers and expert evaluators of highly interactive websites.

The first step was to categorize all the usability problems using a grounded theory approach to allow the categories to emerge themselves. This resulted in four major categories of Physical Presentation, Content, Information Architecture and Interactivity. These, of course, are four major themes of discussion about the web and interactive systems, so these categories are not surprising in themselves, but it is interesting that these major categories emerged. Within each major category, a number of sub-categories emerged, with Interactivity having the largest number of sub-categories. Again, this is perhaps not surprising, as interactivity is the newest aspect of the design of websites, as websites move towards Web2.0 and become more interactive than the informational websites typical of the 1990s and early 2000s. Thus web developers may be less familiar and confident about how to produce these aspects of websites.

The next step was to analyze those problems that users were likely to encounter and experts were likely to miss and vice versa. This revealed some unexpected results. Users were much more likely than experts to encounter security issues, input format problems and poor color contrast, all problem areas that we expected experts to be monitoring for carefully. However, it was less surprising that experts were more likely than users to find problems with consistency of the "look and feel" of the website, unexplained terminology and saliency of key content and elements, as these are areas that experts ought to be monitoring for carefully.

From these analyses, we have proposed a set of heuristics for the development and evaluation of highly interactive websites that are evidence-based, using both the severity and frequency of problems encountered by users. Unfortunately, this yields a rather lengthy set of 21 heuristics (covering 23 sub-categories of our emergent categorization), but grouped into the four major categories. Physical presentation has four heuristics, Content three heuristics, Information Architecture only one heuristic and Interactivity 13 heuristics.

We made a comparison of these new heuristics with Molich and Nielsen's heuristics [17, 18, 20], probably the best known and most widely used of the available heuristic sets.

HEURISTIC	Rationale for inclusion
PHYSICAL PRESENTATION	
1. Make text and interactive elements large and clear enough Default and typically rendered sizes of text and interactive elements should be large enough to be easy to read and manipulate.	Rationale: Frequency of problem for users Frequency: 18 times
2. Make page layout clear Make sure that the layout of information on the page is clear, easy to read and reflects the organization of the material.	Rationale: Frequency Frequency: 18 times
3. Avoid short time-outs and display times Provide time-outs that are long enough for users to complete the task comfortably, and if information is displayed for a limited time, make sure it is long enough for users to read comfortably.	Rationale: Severity of problem for users Median severity rating = 2.25
4. Make key content and elements and changes to them salient Make sure the key content and interactive elements are clearly visible on the page and that changes to the page are clearly indicated.	Rationale: Frequency Frequency: 14 times
CONTENT	
5. Provide relevant and appropriate content Ensure that content is relevant to users' task and that it is appropriately and respectfully worded.	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 21 times
6. Provide sufficient but not excessive content Provide sufficient content (including Help) so that user can complete their task but not excessive amounts of content that they are overwhelmed.	Rationale: Frequency and severity Median severity rating (sufficient) = 2.47 Frequency: 47 times Median severity rating (excessive) = 2.0
7. Provide clear terms, abbreviations, avoid jargon Define all complex terms, jargon and explain abbreviations.	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 19 times
INFORMATION ARCHITECTURE	
8. Provide clear, well-organized information structures Provide clear information structures that organize the content on the page and help users complete their task.	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 30 times
INTERACTIVITY	
9. How and why Provide users with clear explanations of how the interactivity works and why things are happening.	Rationale: Frequency and severity Median severity rating = 2.2 Frequency: 51 times
10. Clear labels and instructions Provide clear labels and instructions for all interactive elements. Follow web conventions for labels and instructions (e.g. use of asterisk for mandatory elements).	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 48 times
11. Avoid duplication/excessive effort by users Do not ask users to provide the same information more than once and do not ask for excessive effort when this could be achieved more efficiently by the system.	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 14 times
12. Make input formats clear and easy Make clear in advance what format of information is required from users. Use input formats that are easy for users, such as words for months rather than numbers.	Rationale: Frequency Frequency: 18 times
13. Provide feedback on user actions and system progress Provide feedback to users on their actions and if a system process will take time, on its progress.	Rationale: Severity Median severity rating = 2.0
14. Make the sequence of interaction logical Make the sequence of interaction logical for users (e.g. users who are native speakers of European languages typically work down a page from top left to bottom right, so provide the Next button at the bottom right).	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 11 times
15. Provide a logical and complete set of options Ensure that any set of options includes all the options users might need and that the set of options will be logical to users.	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 57 times

16. Follow conventions for interaction Unless there is a very particular reason not to, follow web and logical conventions in the interaction (e.g. follow a logical tab order between interactive elements).	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 46 times
17. Provide the interactive functionality users will need and expect Provide all the interactive functionality that users will need to complete their task and that they would expect in the situation (e.g. is a search needed or provided?).	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 73 times
18. Indicate if links go to an external site or to another webpage If a link goes to another website or opens a different type of resource (e.g. PDF document) indicate this in advance.	Rationale: Severity Median severity rating = 2.0
19. Interactive and non-interactive elements should be clearly distinguished Elements which are interactive should be clearly indicated as such, and element which are not interactive should not look interactive.	Rationale: Frequency Frequency: 14 times
20. Group interactive elements clearly and logically Group interactive elements and the labels and text associated with them in ways that make their functions clear.	Rationale: Frequency Frequency: 15 times
21. Provide informative error messages and error recovery Provide error messages that explain the problem in the users' language and ways to recover from errors.	Rationale: Frequency and severity Median severity rating = 2.0 Frequency: 15 times

Table 6. New evidence-based heuristics for designing and evaluating highly interactive websites.

.

9 of the 21 (42.9%) new heuristics do not feature in those heuristics. These are new heuristics #1, #3, #5, #6, #15, #17, #18, #19 and #20. Further, five of the new heuristics share aspects from more than one Molich and Nielsen heuristic. These are heuristics #4, #9, #10, #11 and #12. Finally, seven of the new heuristics map onto one Molich and Nielsen heuristic; however, this mapping is not one-to-one. A number of the new heuristics share a Molich and Nielsen heuristic in common. For example, the Molich and Nielsen heuristic "Match between system and real world" specifically says that designers should "*Follow real-world conventions, making information appear in a natural and logical order*". The new heuristics "Make Page Layout Clear" and "Make sequences of action logical" are two heuristics that are more precise, addressing different aspects of the Molich and Nielsen heuristic as it applies to current websites.

This demonstrates that the new heuristics are both different in coverage from Molich and Nielsen's and different in their organization. This is not a criticism of Molich and Nielsen's work, but reflects the very different nature of current highly interactive websites in comparison to the interfaces of the 1980s from which Molich and Nielsen drew their heuristics. Indeed, the overlap with Molich and Nielsen's heuristics may actually be less obvious for a practicing web developer or evaluator. We found that when working with our large corpus of problems we could map backwards from current problems to Molich and Nielsen's heuristics. However, without this corpus, when we have been conducting evaluations of individual websites, it has been very difficult to map forwards from the Molich and Nielsen heuristics to problems encountered with current websites. For example, the new heuristic "Avoid duplication/excessive effort by users" is equivalent to aspects of both Molich and Nielsen "Aesthetic and minimalist design" and "Recognition rather than recall", but it is not clear that evaluators recognize the kinds of problems grouped under the new heuristic as being exemplars of the two original Molich and Nielsen heuristics.

Further work is needed to establish whether these heuristics are indeed more effective in the development and evaluation of websites, beyond the fact that they are developed from a large corpus of problems. Our own future work will involve using the new set of heuristics in expert evaluation of a further set of highly interactive websites and comparing the results with user evaluation of the same websites. We would predict that using the new heuristics should guide evaluators to the problems that users encounter and yield a higher overlap in problems between user and expert evaluation, thus improving the effectiveness of the expert evaluation. However, it is important that independent researchers also conduct evaluations using these heuristics, in both development and evaluation contexts, and we welcome such studies.

If our evaluation study is successful, further work will explore the generalizability of the heuristics to more diverse websites and other interactive systems. An important question to address is whether it possible to have a general set of heuristics to capture the main usability problems of all interactive systems, or is it the case that the scope of interactive systems is now so broad, that different heuristics are needed for different categories of interactive system? Certainly, the new set of heuristics is already large, and adding more heuristics to cope with a wider range of interactive systems may defeat the purpose of a set of heuristics that is relatively easy to remember and use.

In conclusion, we believe that an evidence-based set of heuristics for highly interactive websites is a useful tool to both developers producing websites and experts evaluating them. In particular, use of these heuristics should improve the effectiveness of expert evaluation of websites.

ACKNOWLEDGMENTS

The authors thank Lucy Buykx, Andre Freire, John Precious and David Swallow for their assistance with collecting and coding of the data for this paper. They also thank all the participants and experts who participated and DirectGov (www.direct.gov.uk) for their funding of the work.

REFERENCES

1. Batra, S. and Bishu, R.R. (2007). Web usability and evaluation: issues and concerns. In N. Aykin (Ed.), *Usability and Internationalization, Part I, HCII 2007* (LNCS 4559). Berlin: Springer-Verlag.
2. Bryant, T. and Charmaz, K. (2007). *The Sage Handbook of Grounded Theory*. London: Sage.
3. Budd, A. (2007). Heuristics for modern web application development. Blogography, January 17. Available at: www.andybudd.com/archives/2007/01/heuristics_for_modern_web_application_development/
4. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 37–46.
5. Cockton, G. and Woolrych, A. (2001). Understanding inspection methods: lessons from an assessment of heuristic evaluation. In A. Blandford, J. Vanderdonckt and P.D. Gray (Eds.), *People and Computers XV*. Berlin: Springer Verlag.
6. Dix, A., Finlay, J., Abowd, G.D. and Beale, R. (2004). *Human-computer interaction* (3rd edition). Harlow, UK: Pearson Prentice Hall.
7. Gray, W.D. and Salzman, M. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human Computer Interaction*, 13(3), 203 – 261.
8. Hertzum, M. and Jacobsen, N.E. (2001). The evaluator effect: a chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13(4), 421 – 443.
9. Hertzum, M., Jacobsen, N.E. and Molich, R. (2002). Usability inspections by groups of specialists: perceived agreement in spite of disparate observations. In *Ext. Abstracts CHI 2002*, ACM Press (2002), 662-663.
10. Hornbaek, K. (2011). Dogmas in the assessment of usability evaluation methods. *Behavior and Information Technology*, 29(1), 97 – 111.
11. Instone, K. (1997). Site usability heuristics for the web. Web Review, October. Available at: http://instone.org/heuristics
12. Jeffries, R., Miller, J.R., Wharton, C. and Uyeda, K. (1991). User interface evaluation in the real world: a comparison of four techniques. *Proc. CHI 1991*, ACM Press (1991), 119-124.
13. Koutsabasis, P., Spyrou, T. and Darzentas, J. (2009). Evaluating usability evaluation methods: criteria, method and a case study. In J. Jacko (Ed.), *Human-Computer Interaction, Part I, HCII 2007* (LNCS 4550). Berlin: Springer Verlag.
14. Lazar, J., Feng, J.H. and Hochheiser, H. (2010). *Research methods in human-computer interaction*. Chichester, UK: Wiley.
15. Lewis, C., Polson, P., Wharton, C., and Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. *Proc. CHI 1990*, ACM Press (1990), 235-242.
16. Lewis, C. and Wharton, C. (1997). Cognitive walkthroughs. In M. Helander, T. K. Landauer and P. Prabhu (Eds.), *Handbook of human–computer interaction (2nd edition)*. Amsterdam: Elsevier.
17. Molich, R. and Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM*, 33(3), 338 – 348.
18. Nielsen, J. (1993). *Usability engineering*. San Diego, CA: Morgan Kaufmann.
19. Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. *Proc. CHI 1994*, ACM Press (1994), 152-158.
20. Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces. *Proc. CHI 1990*, ACM Press (1990), 249-256.
21. Petrie, H. and Buykx, L. (2010). Collaborative Heuristic Evaluation: improving the effectiveness of heuristic evaluation. *Proceedings of UPA 2010 International Conference.* Omnipress. Available at: http://upa.omnibooksonline.com/index.htm
22. Rogers, Y., Sharp, H. and Preece, J. (2011). *Interaction design: beyond human-computer interaction.* Chichester, UK: Wiley.
23. Shneiderman, B. and Plaisant, C. (2005). *Designing the user interface (4th edition).* Boston, MA: Addison Wesley.
24. Strauss, A. and Corbin, J. (1997). *Grounded theory in practice*. London: Sage.
25. Tognazzini, B. (2003). First principles of interaction design. Available at: http://asktog.com/basics/firstPrinciples.htm