Publisher policy allows this work to be made available in this repository. Published in *Library Reviews* by Emerald. The original publication is available at: https://doi.org/10.1108/00242530910997928. This article is deposited under the Creative Commons Attribution Non-commercial International Licence 4.0 (CC BY-NC 4.0). Any reuse is allowed in accordance with the terms outlined by the licence (https://creativecommons.org/licenses/by-nc/4.0/). To reuse the AAM for commercial purposes, permission should be sought by contacting permissions@emeraldinsight.com.

Evaluating the usability *and* usefulness of a digital library

Steven Buchanan* and Adeola Salako

Department of Computer and Information Sciences, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, UK

*** Correspondence to steven.buchanan@cis.strath.ac.uk**

Abstract

Purpose System usability and system usefulness are interdependent properties of system interaction, which in combination, determine system satisfaction and usage. Often approached separately, or in the case of digital libraries, often focused upon usability, there is emerging consensus among the research community for their unified treatment and research attention. However, a key challenge is to identify, both respectively and relatively, *what* to measure and *how*, compounded by concerns regarding common understanding of usability measures, and associated calls for more valid and complete measures within integrated and comprehensive models.

Design/methodology/approach Identified key usability and usefulness attributes and associated measures, compiled an integrated measurement framework, identified a suitable methodological approach for application of the framework, and conducted a pilot study on an interactive search system developed by a Health Service as part of their e-library service.

Findings Effectiveness, efficiency, aesthetic appearance, terminology, navigation, and learnability are key attributes of system usability; and relevance, reliability, and currency key attributes of system usefulness. There are shared aspects to several of these attributes, but each is also sufficiently unique to preserve its respective validity. They can be combined as part of a multi-method approach to system evaluation.

Originality/value Provides an integrated measurement framework, derived from the goal, question, metric paradigm, which provides a relatively comprehensive and representative set of system usability and system usefulness attributes and associated measures, which could be adapted and further refined on a case-by-case basis.

Research limitations/implications Pilot study has demonstrated that usability and usefulness can be readily combined, and that questionnaire and observation are valid multi-method approaches, but further research is called for under a variety of conditions, with further combinations of methods, and larger samples.

Keywords System usability, System usefulness, Digital libraries.

Paper type Research paper

1. Introduction

System usability and system usefulness are related properties of system interaction (Tsakonas & Papatheodorou, 2006), which in combination, determine system satisfaction and usage. While usability evaluations might lead to more usable systems, it is argued that without consideration of usefulness, systems could prove to be effectively designed, but functionally useless (Greenberg and Buxton, 2008). Further, consideration of usefulness not only facilitates use, but also improvement and innovation (Greenberg and Buxton, 2008).

Often approached separately (Dicks, 2002), or with emphasis upon usability (for example, Xie [2008] reports that the majority of digital library evaluation studies are usability studies), there is emerging consensus among the research community for their unified treatment and research attention (Tsakonas and Papatheodorou, 2008). However, a key challenge is to identify, both respectively and relatively, *what* to measure and *how*, compounded by concerns regarding common understanding of usability measures, and associated calls for more valid and complete measures within integrated and comprehensive models (Hornbak, 2006; Abran et al. 2003).

With particular attention to identification of respective attributes, associated measures, and the relationship between, this study sought to identify how usability evaluation might be extended to usefulness, and to then conduct a pilot test of an appropriate approach to their combined evaluation. The test case was a recently launched clinical decisions portal developed by a Health Service as part of their e-library service, which was developed to provide clinicians with direct access to clinical evidence and best practice recommendations to support decision-making at point of care, and to support clinicians' ongoing learning and professional development.

2. Key attributes and associated measures

Usability is concerned with aspects of human computer interaction and in particular, the user interface. In contrast, usefulness is concerned with whether or not the system supports user activity (Burns et al, 1997; Kushniruk and Patel, 2004). The distinction is akin to one of form versus function.

2.1. Usability

Several usability attributes have been variously proposed to guide measurement. With respect to associated standards, ISO 9126-1 specifies understandability, learnability, operatability, and attractiveness (and extending to usability compliance), while ISO 9241-11 specifies effectiveness, efficiency, and satisfaction. Within the research community, Nielson (1993) notably proposed learnability, efficiency, memorability, errors and satisfaction, while in more recent studies, Abran et al (2003) proposes effectiveness, efficiency, satisfaction, security, and learnability, and Tsakonas & Papatheodorou (2006) learnability, ease of use, aesthetic appearance, navigation, and terminology. Attributes may differ across standards and respective authors, but there is noted association in their interpretation. For example, ease of use has been associated with efficiency (Dicks, 2002) and operability (Hanson and Castleman, 2006), errors with effectiveness (Folmer & Bosch, 2004), and terminology with both understandability and memorability (Yushiana and Rani, 2007). It could also be argued that aesthetic appearance is similar to attractiveness, which has been further associated with satisfaction (Folmer & Bosch, 2004).

While the above review is not exhaustive, it nonetheless references key standards, and provides an indication of current opinion regarding key attributes, and the relationships between. With this in mind, we elected to proceed with the following attributes: effectiveness, efficiency, aesthetic appearance, terminology, navigation, and learnability, which were then defined in more depth, with particular attention to respective and relative validity.

Effectiveness

Effectiveness is concerned with task completion in relation to user goals, in particular success rates. According to ISO 9241, related attributes are accuracy and completeness. Typically measured by task completion (e.g. information required located), this can extend to percentage of tasks completed, percentage of tasks completed per unit of time, and ratio on failure handling (Abran et al., 2003).

Frokjaer et al. (2000) caution that effectiveness is often wrongly omitted from usability studies under the mistaken belief that there is a strong correlation with efficiency. Providing evidence to the contrary, they argue that efficiency and effectiveness should be considered independent aspects of usability, the former concerned with effort, the latter with outcome. Placing emphasis on the importance of outcome, Frokjaer et al. propose quality of solution as the primary indicator of effectiveness (an aspect of effectiveness which we believe is better considered when usability is extended to usefulness [see Section 2.2]).

Efficiency

Efficiency is concerned with task completion in relation to user productivity, in particular time expended (Dicks, 2002). Task completion time is considered a valid measure (Petrelli, 2007), but can extend to error percentage, time spent on errors, repetitions' number of failed commands (e.g. responded without undue delay or error), documentation or help's use frequency, number of good and bad characteristics recalled by users, and number of available commands not called upon (Abran et al., 2003).

It has been suggested that task completion time is not suitable for web-based systems as external factors such as connection speed and network traffic could adversely affect the time taken to display a web page or process a request (Benbunan-Fich, 2001), and that task completion time defeats the purpose of those web based systems which encourage browsing (Smith, 1996); however, while these are valid considerations (the latter particularly so given the iterative nature of information seeking), a study by Yu and Kaufman (2007) observed that users (in this case physicians) spent on average two minutes or less seeking an answer to a question and if a search took longer, it was likely to be abandoned, suggesting that time expended is a valid consideration.

Aesthetic appearance

Aesthetics refers to the consistency and appropriateness of the system interface design, in particular layout, colours, fonts, and graphic properties (Kirmani and Rajasekaran, 2005). Aesthetics, and the associated concept of attractiveness, have been shown to be strongly correlated to user perceptions of system usability (Tractinsky et al, 2000; De Angeli et al, 2006; Dillon, 2001) in a relationship referred to as the 'halo effect', a reference to the way human perception of beauty causes a favourable disposition towards the object in question (Hartmann, 2006; De Angeli et al, 2006), which can extend to perceptions of credibility (Wathen and Burkell, 2002).

Tractinsky (1997) has argued for aesthetics to include considerations of how well they facilitate information processing (e.g. used appropriately). Corroborating this view are results from a study by Hu et al (1999) where it was observed that graphical interfaces in IR systems which incorporated the use of size, distance and color in pointing out relevant items in response to a user's search query were more effective than designs based on only one of these visual properties.

Navigation

Navigation refers to the ease with which the user can traverse the interface using the navigation tools available to them (bars, icons, menus, colour/typographic codings etc.), and at any point in time, how aware they are of their current location. Location awareness is a key aspect of navigation (Aiita et al, 2008; Hassan and Li, 2005), as disorientation can lead to cognitive overload (Pearson et al, 2007). The disorientation termed 'lostness' which follows, is an occurrence of concern as it has been shown to reduce use of web based systems (Otter and Johnson, 2000; Smith, 1996).

Navigation is typically measured in terms of the time and steps required to obtain desired results, and how well they can control what they are doing and where they are (Flavian et al, 2005).

Terminology

Terminology considers how well the user can comprehend the terms and phrases used to describe functions or content within the interface (Tsakonas and Papatheodorou, 2006), and the consistency of terms used and how logically they have been placed (Aiita et al, 2008). Communication between two entities (in this case the system and user) can only take place when they share a common language (Yushiana and Rani, 2007), however despite the recognition of this fact, it has been observed that system developers often use jargon or designer centred language rather than user centred language when designing system interfaces (Hartson et al, 2004), which then adversely impacts navigation and retrieval (Yushiana and Rani, 2007). Unfamiliar terminology is often attributed to the difficulties system developers face in finding a common language to use in term descriptions for web interfaces, particularly where users come from diverse backgrounds (Aiita et al, 2008). Dzida (1995) recommends that self-descriptive (e.g. logical) explanations derived from the user's task domain be used, particularly where tasks to be executed on the system require user guidance.

Learnability

Learnability refers to the capability of the system to enable users to feel that they can productively use the system right away and quickly learn new functions (Seffah et al, 2006). It is often considered the most fundamental aspect of usability, since learning how to use the system is the first user experience (Nielsen, 1993). It evaluates how easily and effectively the user learns to accomplish tasks, and can be extended to include the contribution of help documentation to the learning process (Glosiene and Manzhukh, 2005). It can also consider how easy it is for infrequent users to relearn the system after periods of inactivity (Rubin, 1994).

Folmer and Bosch (2004) suggest that time taken to learn tasks using the system or number of errors made while performing such tasks are valid objective measures of learnability (as opposed to the more subjective measures above), but note that these should be defined and considered relative to each type of interaction and user.

2.2. Usefulness

The content and services offered by a system, and how closely they meet user requirements, are considered key aspects of system usefulness (Hartmann, 2006; Savolainen, 2008). Similar to usability (albeit to a lesser degree), various attributes have been proposed to guide measurement. For example, Yang et al (2005) proposes value, reliability, currency, and accuracy, while Tsakonas & Papatheodorou (2008) propose relevance, format, reliability, level, and coverage. Also similar to usability, there is association in their interpretation. For example Yang et al. (2005), have themselves associated value with relevance, and accuracy with reliability, while Vakkari and Hakala (2000) have associated level with relevance, and Xie (2006) coverage with reliability. In consideration of the above, we selected relevance, reliability, and currency. Similar to usability, we felt that this was a reflective selection. Each is discussed in turn below.

Relevance

Relevance, considered to be one of the most fundamental aspects of information retrieval (Tombros et al, 2004), is a multi-dimensional concept, as it relates to content, which can be considered objective, and also relates to the particular experience and needs of the user, which can be considered subjective (Thornley and Gibb, 2007). According to Barry and Schambler (1998) relevance considers both the users' (cognitive) knowledge and (subjective) perceptions, is situational (influenced by the information problem), complex and multidimensional, and although dynamic and constantly changing, is also systematic, observable and measurable at a single point in time.

Within the context of system usefulness, relevance is associated with how well the system enables the accomplishments of user tasks and in particular, how well information retrieved contributed to the user requirement. Associated attributes are pertinence and utility (Greisdorf, 2002). Topicality, which denotes the extent to which system output matches the user provided search word or specifications (Hu et al, 1999), is also considered a key measurement of relevance (Borlund and Ingwersen, 1997; Tsakonas and Papatheodorou, 2006; Xie, 2006).

Reliability

Reliability refers to the accuracy, dependability, and consistency of information (Yang et al, 2005), and is associated with credibility (Tsakonas and Papatheodorou, 2006), a complex cognitive process by which information is filtered and selected (Liu, 2004). Credibility will to a large extent determine whether or not the resource is accepted and put to further use (Burgoon et al, 2000).

Wathen and Burkell (2002), demonstrating the complex interrelationships between usability and usefulness, propose that there are three stages of user interaction which establish credibility: firstly, the 'surface' level based upon aspects of usability (appearance, interface design, organization of information); secondly the 'message' level based upon credibility of source (expertise/competence, trustworthiness, credentials), and credibility of message (content, relevance, currency, accuracy, tailoring); and thirdly the 'content' level, based upon the user's cognitive state (knowledge, motivation).

Currency

Currency considers the extent to which the information is sufficiently up-to-date for the task it is to be used for (Pipino et al., 2002). Although currency is relative to domain and task, users generally attach high value to current information (Xie, 2006), with information retrieved from out-of-date collections no longer considered accurate. However, Gonçalves et al (2006) note that information may not always be up-to-date, but may remain valid based upon overall importance with the community of interest (Wang et al [1995] refer to this as the 'volatility' of the information). As a consequence, Goncalves et al argue that not only is creation date a valid indicator of currency, but also time of last citation.

2.3. A Measurement Framework

There are interdependent relationships and associated overlap between the selected attributes, but in our opinion, each also possesses sufficient uniqueness of purpose to preserve its respective validity. These attributes and associated key measures are summarised in Table 1 in a manner derived from the goal, question, metric paradigm, which promotes an analysis driven measurement approach (Kan, 2003).

We acknowledge that there is a degree of subjective interpretation in our selection of these attributes and associated measures, and that in everyday use there are influencing factors to consider such as user, task, and environment (Barry and Schambler, 1998; Frokjaer et al., 2000; Abran et al., 2003); however we feel that this selection provides a relatively comprehensive and representative set, which could be adapted and further refined on a case by case basis. Importantly, it is not proposed as an amendment to existing standards, but as an accompaniment.

Table 1. Usability and Usefulness: a measurement framework.

GOAL (Improve…)	QUESTION (Asks if…)	METRIC (measures…)
Effectiveness	Information required was located	Tasks completed
Efficiency	The system responded quickly to the task (without delay or error)	Time to complete
Aesthetic Appearance	Text type and font size are engaging and readable	Attractiveness
	Colours, graphics, and icons have been used appropriately	Appropriateness
Terminology	The terms used to label the menu functions are understandable	Comprehension
	The menu functions are logically related	Consistency
Navigation	Orientation is straightforward	Steps to complete
Learnability	Steps required to complete tasks were understandable	Repetition failed commands
Relevance	Information retrieved reflected the query	Relevant results
	Information retrieved contributed to the requirement	Utility
Reliability	Information retrieved was from a credible source	Credibility
Currency	Information retrieved is current	Creation Date
	Information retrieved is valid	Last Citation

We next considered an appropriate approach to combined evaluation.

3. Methodological Approach

Usability evaluation can be both formative and summative, and is commonly conducted by inspection and/or test, the former without involvement of the user, the latter typically with. Inspection methods include heuristic evaluation, cognitive walkthrough, and action analysis, while test methods include questionnaire, thinking aloud, and field observation (Holzinger, 2005). In contrast to usability, usefulness is much more dependent upon user involvement. It can be considered during formative stages of system design (based on user input/statement of requirement or functioning prototype/simulation), but evaluation is dependent upon user interaction, within context, and preferably under live conditions. As a consequence, usefulness evaluation is commonly conducted by field observation.

Our pilot study was summative and test-oriented, being conducted on a recently deployed system, and at client request, focused upon ascertaining user satisfaction. When considering an appropriate approach we noted that there is general consensus that no single evaluation technique yields the best results (Karat et al, 1992; Lavery et al, 1997; Molich and Jeffries, 2003), and that multi-method approaches accommodate organizational constraints, enable wider user involvement, and facilitate validation (Glosiene and Manzhukh, 2005). We also noted Holzinger's (2005) recommendation that, wherever possible, indirect evaluation methods are supported by direct evaluations to allow comparison of stated versus actual behavior. This led us to questionnaire and field observation, both of which are proven evaluation techniques that have been successfully combined in previous studies (Aborg et al, 2002). We also considered thinking aloud, but while we acknowledged that this could provide valuable insight into the users' mental model and interaction with the system, we were also concerned that this might not necessarily be a true representation of users' real world perceptions (Holzinger, 2005; Aitta et al, 2008).

An 18-point electronic questionnaire was developed, with questions drawn from the previously identified key measures (see Table 1). Participants were instructed to identify an information need related to patient care, use the system to retrieve the information, and then complete the questionnaire. Each question had associated end points ranging from (1) strongly disagree to (4) strongly agree, a scale derived from the Computer System Usability Questionnaire (Lewis, 1993). Each question included supporting definition and provided opportunity for additional comment. The questions in Table 1 were preceded by three demographic questions (age [bands], organizational role, and avg. time spent online [per week])[1], and followed by three general questions (system has all expected functionality; overall I am satisfied with the system; I would use the system again) and one final specific question, which asked if there was one thing that could be done to improve the system, what would it be?

For the observation based tests, tasks to be performed were set by participants based upon a hypothetical or real medical case, providing a more realistic test-case scenario framed within an operational context (Borlund, 2000; Hornbak, 2005; Granic, 2008), and preserving the ecological validity of the study (Haynes et al, 2004; Petrelli, 2007; Gordon and Pathak, 1999). This would also help ensure that the task was appropriate to the level of experience of the participants (Rubin, 1994). Tasks were conducted on location within the user environment, but not in the presence of patients.

A challenge with observation is how to effectively observe in a non-intrusive way. One approach is to attempt to discretely video record participants completing tasks, but this may prove difficult in practice, as both interface and user must be in detailed and close shot to facilitate observation. Video recording can also be time consuming, both in setup and later analysis. Holzinger (2005) considers video rarely necessary, arguing that key observations will be obvious to the observer, while Nielson (2000) considers it to be an unnecessary overhead, which more importantly, can intimidate users. In consideration of this we elected to observe without recording equipment, considering this to be less intrusive. For related reasons neither would the observer respond to any unsolicited participant comment during observation, reasoning that discussion, although potentially valuable to the observer, might interrupt or influence the user's cognitive process. The observer would note unsolicited comments, but would (politely) not enter discussion until the exercise was completed. Observation and noted comments were recorded and coded against associated attribute.

Finally, and in accordance with an additional client requirement to benchmark the system, participants were asked (post-observation) to compare the system's performance with an alternative commercially available system, an act of comparative analysis that would evaluate how well each system supported user tasks, and potentially lead to improvements based upon consideration of their respective strengths and weaknesses (Ahmed et al, 2006, Hassan and Li, 2005). Participants were asked to repeat their tasks with the second system, and to then answer the questions in Table 1, but with end points now ranging from worse (1), to similar (2), to better (3), to much better (4).

[1] Participants in the observation-based tests were also asked to provide this information.

Volunteer participation (questionnaire and observation) was sought via the Health Service librarian network, eHealth and clinical education leads, and associated electronic distribution lists.

4. Results

4.1 Questionnaire

30 clinicians responded to the questionnaire with one incomplete return, which was discounted. Of the 29 completed questionnaires, approximately half of respondents occupied nursing, midwifery, and hospital medicine roles, with the remainder evenly distributed across general practice and the allied health professions. The age range was from 20 to 45+ with approximately half aged over 45. With respect to time spent per week online (work-related), one respondent (3.4%) spent no time online, twenty-four respondents (82.8%) spent between 5-9 hours each week online, and the remaining four (13.8%) spent 10+ hours per week online.

Questionnaire results were positive overall (see Table 2); however, some dissatisfaction was noted through additional comment (and reflected in mean scores), in particular with regard to aspects of efficiency, terminology, navigation, and relevance. With regard to efficiency, some respondents commented that response time was too slow, and that too much effort was required to (repeatedly) enter passwords, and filter information; with regard to terminology, that some terms were difficult to understand and some labeling obscure; with regard to navigation, that the system was too complex, with some reporting having reached dead ends while seeking information; and with regard to relevance, some reported irrelevant results. Nonetheless, in a reflection of the positive ratings overall, when asked if they would use the system again, 25 (86.20%) responded yes.

Table 2 Questionnaire Results

	1 Strongly Disagree n (%)	2 Disagree n (%)	3 Agree n (%)	4 Strongly Agree n (%)	Median
Information required was located	1 (3.45)	5 (17.24)	18 (62.07)	5 (17.24)	3
The system responded quickly to the task (without undue delay or error)	1 (3.45)	3 (10.34)	18 (62.07)	7 (24.14)	3
Text type and font size are consistent and readable	1 (3.45)	1 (3.45)	23 (79.31)	4 (13.79)	3
Colours, graphics, and icons have been used appropriately	1 (3.45)	2 (6.90)	23 (79.31)	3 (10.34)	3
The terms used to label the menu functions are understandable	1 (3.45)	4 (13.79)	21 (72.41)	3 (10.34)	3
The menu functions are logically related	1 (3.45)	4 (13.79)	20 (68.97)	4 (13.79)	3
Orientation is straightforward	2 (6.90)	6 (20.69)	16 (55.17)	5 (17.24)	3
Steps required to complete tasks were understandable	1 (3.45)	4 (13.79)	21 (72.41)	3 (10.34)	3
Information retrieved reflected the query	2 (6.90)	3 (10.34)	14 (48.28)	10 (34.48)	3
Information retrieved contributed to the requirement	3 (10.34)	3 (10.34)	22 (75.86)	1 (3.45)	3
Information retrieved was from a credible source	2 (6.90)	2 (6.90)	17 (58.62)	8 (27.59)	3
Information retrieved is current	1 (3.45)	4 (13.79)	19 (65.52)	5 (17.24)	3
Information retrieved is valid	1 (3.45)	2 (6.90)	20 (68.97)	6 (20.69)	3
System has all expected functionality	1 (3.45)	5 (17.24)	21 (72.41)	2 (6.90)	3
Overall I am satisfied with the system	1 (3.45)	5 (17.24)	20 (68.97)	3 (10.34)	3

The final question, which asked respondents to identify one thing that might be done to improve the system, also proved informative, with five clear recommendations emerging from grouped comments: increased use of colour to guide interaction; provision of an online guide, particularly for constructing search queries;

single log-on; retrieved documents presented in order of relevance; and increased awareness of the system among staff.

4.2 Observation

Seven clinicians volunteered to participate, but unfortunately three had to withdraw at short notice due to unavoidable engagements. Of the four who participated, two were general practitioners, one a consultant surgeon, and one a podiatrist. Three of the participants spent up to 5 hours per week online, and one 5-9 hours online. All were aged over 45.

Under observation participants appeared frustrated with repeat log-on to various sites, were observed to experience some difficulty in constructing their search queries (two of the four participants, when adopting more unstructured natural language, failed to obtain relevant results); and relied on the browser back button to navigate back to the portal homepage, which in one instance led to the user becoming completely disoriented.

Participants later commented (post-observation) that navigation was not straightforward, that the system appeared slow (although acknowledged as possibly network related), and that terminology was not always self-explanatory. Aesthetics were praised, but not considered key. One participant commented on the irrelevance of the documents retrieved, suggesting that too many broad terms might have been used to index the documents. Participants suggested that the system would benefit from single log-on, summaries of documents retrieved, sample queries, and increased error tolerance (e.g. for misspelled terms). Comparisons were repeatedly made with Google, suggesting that this was a favored search engine amongst participants. Notably, the overall purpose of the system was not self apparent to participants, with two users questioning its function in relation to the existing e-library service.

After repeating tasks with the commercial alternative (three of the four participants obtaining relevant results), participants considered the first (in-house) system (median score in brackets) better with regard to aesthetics (3.00) and currency (3.00), similar with regard to efficiency (2.00), relevance (2.00), and reliability (2.00), but worse with regard to terminology (1.00), navigation (1.50), and learnability (1.50). Overall, three of the four participants indicated a preference for the second (commercial) system, citing simplicity (less steps) and speed of retrieval as the deciding factors (considered key by participants as within live clinical settings a consultation would typically last no more than ten minutes). The fourth participant had no preference.

5. Discussion

Similar issues were raised and improvements suggested by both questionnaire and observation participants, particularly with regard to aspects of efficiency, terminology, navigation, and relevance; however, it is notable that while questionnaire results were in general positive, the results obtained via observation were less so, with participants more critical of the system.

A possible explanation for the more positive ratings obtained from the questionnaire returns is provided by Kelly et al (2008), who argue that when completing questionnaires, users have a tendency to inflate system ratings, even when the system violates basic usability principles. The reasons for this, it is argued, are threefold (Kelly et al, 2008): users can tend to agree with attitude statements when presented to them; often assume that there is a demand for them to behave in a particular way; and can view success or failure to complete a system task as a reflection of their own abilities rather than as a reflection of the system's abilities. A further consideration is whether or not the user accurately reflected on the true experience, particularly if completing a questionnaire at a later point (Webster and Williams, 2005).

While these are valid considerations, the questionnaire nonetheless allowed us to survey a larger number of participants than would have been possible with observation alone, a reason why, despite its limitations, the questionnaire remains widely used (Folmer and Bosch, 2004). It should also be noted that valuable user comments and recommendations were obtained from the questionnaire, which the observation-based tests further supported. In our opinion the questionnaire remains valid, but in line with current thinking, preferably as part of a multi-method approach, as our results have reminded us of the importance of paying as much attention to what users do, as to what they say (Webster and Williams, 2005).

The observation-based tests in particular demonstrated the inter-related nature of usability and usefulness, and the benefits of combined evaluation. For example, during observation two participants failed to obtain relevant results and were observed to quickly lose interest in the system (supporting Rubin's [1994] assertion that even if a system is usable, it will only be used if it is also useful), yet later made positive comments regarding aspects of usability. Without observation, these comments might have been misleading.

Benchmarking also proved valuable, encouraging users to compare functionality and, as anticipated, identify respective strengths and weaknesses (Hassan and Li, 2005). It is possible, that if we had asked questionnaire respondents to also undertake the benchmarking exercise ratings might have been closer, as comparison might have encouraged further critique. However this would have significantly increased the time to complete and might have influenced participation.

Both questionnaire respondents and observation participants appeared to readily accept and intuitively understand the presented usability and usefulness attributes and measures, with no contradictory ratings or comments returned, nor confusion observed. The evidence from this study suggests that they are readily combinable, supporting Tsakonas and Papatheodorou's (2008) findings.

However two limitations to consider with our pilot was firstly the lack of objective measurement, and secondly the low number of observation participants. Evaluation of usability (and usefulness) is considered to be to a large degree subjective, being related to users' perception of the interface, interaction or outcome (Folmer and Bosch, 2004; Petrelli, 2007); however, it would have been good to have incorporated objective and quantifiable measures (acknowledging that there are also challenges to consider in distinguishing between and empirically comparing subjective and objective measures of usability [Hornbak, 2005]). Ultimately the final evaluation design was influenced by the requirements and constraints of the participating organization (rather than the author's research question), making it difficult to justify more labour intensive and time-consuming quantitative measurement. One benefit however was that the resulting satisfaction oriented questionnaire, being kept relatively simple, encouraged completion (Liu, 2004).

With regard to the low number of observation participants, we had sought 8-10, which is considered an acceptable 'small' sample (20 being an approximate upper end), particularly were participants are representative users (Kushnniruk, 2004); however, only seven volunteers came forward, and three had to withdraw at the last minute due to conflicting engagements. In our support, Dicks (2002) notes that few usability/usefulness tests are applied to large 'statistically acceptable' samples due to resource and time constraints, and argues that although limited testing might not verify with absolute certainty, it can still provide results of value. We would support this point, as although four participants was less than desirable, valuable observations were still made, and user feedback solicited.

Finally, circumstances dictated a 'snapshot' evaluation, but there are benefits to more longitudinal approaches, particularly for evaluating effectiveness and learnability (Hornbak, 2006), but perhaps more importantly, in relation to relevance. For example, with regard to the clinical decisions portal, an extended study might observe users arriving at a response to a real medical case within a live clinical setting (with patients), observe the learning cycle through repeat observation, and evaluate the contribution of retrieved information to the diagnosis (potentially extending to diagnosis success rates).

6. Conclusion

This study sought to address a key challenge to adopting a unified approach to the evaluation of system usability and usefulness, which was to identify, both respectively and relatively, *what* to measure and *how*; further compounded by concerns regarding common understanding of measures, and associated calls for more valid and complete measures within integrated and comprehensive models.

With regard to *what* to measure, effectiveness, efficiency, aesthetic appearance, terminology, navigation, and learnability have been identified as key attributes of system usability; and relevance, reliability, and currency identified as key attributes of system usefulness. There are shared aspects to each of these attributes, but each is also sufficiently unique to preserve its respective validity, as illustrated by the integrated measurement framework, derived from the goal, question, metric paradigm (see Table 1). This framework is not intended as an alternative to existing standards, but as an accompaniment (providing an integrated and comprehensive model) to guide common understanding of usability and usefulness attributes, and their associated measures. We expect that individual attributes and associated measures will be adapted on a case-by-case basis, and that the framework will be further refined.

With regard to *how* to measure, the pilot study has demonstrated that usability and usefulness can be readily combined, and that questionnaire and observation are valid multi-method approaches, but further research is called for under a variety of conditions, with further combinations of methods, and larger samples.

Referring back to our introduction, usability and usefulness are not just *related* properties, but *dependent* properties of system satisfaction and usage, which with few exceptions, should be jointly considered and evaluated.

References

Aborg, C., Sandblad, B., Gulliksen, J. and Lif, M. (2002). Integrating work environment considerations into usability evaluation methods- the ADA approach. *Interacting with Computers* 15(3) 453-471.

Abran, A., Khelifi, A. and Suryn, W. (2003). Usability meanings and interpretations in ISO standards. *Software Quality Journal* 11(4) 325-338.

Ahmed, S.M.Z., McKnight, C. and Oppenheim, C. (2006). A user-centred design and evaluation of IR interfaces. *Journal of Librarianship and Information Science* 38(3) 157-172.

Aitta, M., Kaleva, S. and Kortelainen, T. (2008). Heuristic evaluation applied to library web services. *New Library World* 109(1-2) 30-43.

Barry, C.L. and Schamber, L. (1998). Users' criteria for relevance evaluation: a cross-situational comparison. *Information Processing and Management* 34(2-3) 219-236.

Benbunan-Fich, R. (2001). Using protocol analysis to evaluate the usability of a commercial web site. *Information and Management* 39(2) 151-163.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *The Journal of Documentation* 53(1) 71-90.

Borlund, P. and Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *The Journal of Documentation* 53(3) 225-250.

Burgoon, J.K., Bonito, J.A.,Bengtsson, B., Cederberg, C., Lundeberg, M. and Allspach, L. (2000). Interactivity in human-computer interaction: a study of credibility, understanding, and influence. *Computers in Human Behaviour* 16(6) 553-574.

Burns, C.M., Vicente, K.J, Christoffersen, K. and Pawlak, W.S. (1997). Towards viable, useful and usable human factors design guidance. *Applied Ergonomics* 28(5-6) 311-322.

De Angeli, A., Sutcliffe, A. and Hartmann, J. (2006). Interaction, usability and aesthetics; what influences users' preferences? In*: Proceedings of the sixth conference on Designing Interactive Systems* (ACM, Pennsylvanaia, 2006) 271-280.

Dicks, R.S. (2002). Mis-Usability: on the uses and misuses of usability testing. In: *Proceedings of the 20th Annual International Conference on Computer Documentation* (ACM, Toronto, 2002) 26-30.

Dillon, A. (2001). Beyond usability: process, outcome and affect in human computer interactions. *Canadian Journal of Information Science* 26 (4) 57-69.

Dzida, W. (1995). Standards for user interfaces. *Computer Standards and Interfaces* 17(1) 89-96.

Flavian, C., Guinaliu, M. and Gurrea, R. (2005). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information and Management* 43(1) 1-14.

Folmer, E. and Bosch, J. (2004). Architecting for usability: a survey. *The Journal of Systems and Software*.70 (1-2) 61-78.

Frokjaer, E., Hertzum, M., and Hornbak, K. (2000) Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: *Proceedings of the SIGCHI conference on Human Factors in computing systems* (ACM Press, The Hague, 2000) 345-352.

Glosiene, A. and Manzhukh, Z. (2005). Towards a usability framework for memory institutions. *New Library World* 106(7-8) 303-319.

Gonçalves, M.A., Moreira, B.L., Fox, E.A. and Watson, L.T. (2006). What is a good digital library?-A quality model for digital libraries. *Information Processing and Management* 43(5) 1416-1437.

Gordon, M. and Pathak, P. (1999). Finding information on the world wide web: the retrieval effectiveness of search engines. *Information Processing and Management* 35(2) 141-180.

Granic, A. (2008). Experience with usability evaluation of e-learning systems. *Universal Access in the Information Society* 7(4) 209-221.

Greenberg, S. and Buxton, B.(2008). Usability evaluation considered harmful (some of the time). In: *Proceedings of the twenty-sixth annual SIGCHI conference on human factors in computing systems* (ACM, Florence, 2008) 111-120.

Greisdorf, H.(2002). Relevance thresholds: a multi-stage predictive model of how users evaluate information. *Information processing and Management* 39(3) 403-423.

Hanson, K. and Castleman, W. (2006).Tracking ease-of-use metrics: a tried and true method for driving adoption of UCD in different corporate cultures. In: Sherman,P.(ed). *Usability success stories: how organizations improve by making easier-to-use software and websites.* England: Gower.

Hartmann, J. (2006). Assessing the attractiveness of interactive systems. In: *CHI '06 extended abstracts on Human factors in Computing Systems* (ACM, Montreal, 2006) 1755-1756.

Hartson, H.R., Shivakumar, P. and Perez-Quinones, M.A. (2004). Usability inspection of digital libraries: a case study. *International Journal of Digital Libraries.*4 (2004), p108-123.

Hassan, S. and Li, F. (2005). Evaluating the usability and content usefulness of web sites: a benchmarking approach. *Journal of Electronic Commerce in Organizations* 3(2) 48-64.

Haynes, S.R., Purao, S. and Skattebo, A.L. (2004). Situating evaluation in scenarios of use. In: *Proceedings of the 2004 conference on Computer supported cooperative work* (ACM, Chicago, 2004) 92-101.

Holzinger, A.(2005). Usability engineering methods of software developers. *Communications of the ACM* 48(1) 71-74.

Hornbak, K. (2006). Current practice in measuring usability: challenges to usability studies and research. *International Journal of Human-Computer Studies* 64(2) 79-102.

Hu, P.J., Ma,P. and Chau, P.Y.K. (1999). Evaluation of user interface designs for information retrieval systems: a computer-based experiment. *Decision Support Systems* 27(1-2) 125-143.

Kan, S.H. (2003) *Metrics and Models in Software Quality Engineering*. Boston: Pearson Education.

Karat, C.M., Campbell, R. and Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In: *Proceedings of the SIGCHI Conference in Human Factors in Computing Systems* (ACM, Monterey, 1992) 397-404.

Kelly, D., Harper, D.J. and Landau, B.(2008). Questionnaire mode effects in interactive information retrieval experiments. *Information Processing and Management* 44(1) p122-141.

Kirmani, S. and Rajasekaran, S. (2005). Heuristic evaluation quality score (HEQS): a measure of heuristic evaluation skills. *Journal of Usability Studies* 2(2) 61-75.

Kushniruk, A.W. and Patel, V.L. (2004). Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of Biomedical Informatics* 37(1) 56-76.

Lavery, D., Cockton, G. and Atkinson, M.P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour and Information Technology* 16(4) 246-266.

Lewis, R.J. (1993). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction* 7(1) 57-78.

Liu, Z. (2004). Perceptions of credibility of scholarly information on the web. *Information Processing and Management* 40(6) 1027-1038.

Molich, R. and Jeffries, R. (2003). Comparative expert reviews. In: CHI ’03 extended abstracts in Human factors in computing systems (ACM, Florida, 2003) 1060-1061

Nielsen, J. (1993). *Usability Engineering.* San Francisco CA: Morgan Kaufmann.

Nielson, J. (2000). *Designing Web Usability*. Indianapolis: New Riders.

Otter, M. and Johnson, H. (2000). Lost in hyperspace: metrics and mental models. *Interacting with Computers* 13(1) 1-40.

Pearson, J.M., Pearson, A. and Green, D. (2007). Determining the importance of key criteria in web usability. *Management Research News* 30(11) 816-828.

Petrelli, D. (2008). On the role of user-centred evaluation in the advancement of interactive information retrieval. *Information Processing and Management* 44(1) 22-38.

Pipino, L.L., Lee.Y.W. and Wang, R.Y. (2002). Data quality assessment. *Communications of the ACM* 45(4) 211-218.

Rubin, J. (1994). *Handbook of usability testing: how to plan, design, and conduct effective tests.* Canada: Wiley.

Savolainen, R. (2008). Source preferences in the context of seeking problem-specific information. *Information Processing and Management* 44(1) 274-293.

Seffah, A., Donyaee, M., Kline, R.B. and Padda, H.K. (2006). Usability measurement and metrics: a consolidated model. *Software Quality Journal* 14(2) 159-178.

Smith, P.A. (1996). Towards a practical measure of hypertext usability. *Interacting with Computers* 8(4) 365-381.

Thornley, C. and Gibb, F. (2007). A dialectical approach to information retrieval. *Journal of Documentation* 63(5) 755-764.

Tombros, A., Ruthven, I., and Jose, J.M. (2004) How users assess web pages for information-seeking. *Journal of the American Society for Information Science and Technology* 56(4) 327-344.

Tractinsky, N. (1997). Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In: *CHI 97, Atlanta, USA, 22-27 March, 1997.* 116.

Tractinsky, N. , Katz, A.S. and Ikar, D. (2000). What is beautiful is usable? *Interacting with Computers* 13(2) 127-145.

Tsakonas, G. and Papatheodorou, C. (2006). Analyzing and evaluating usefulness and usability in electronic information services. *Journal of Information Science* 32(5) 400-419.

Tsakonas, G. and Papatheodorou, C. (2008). Exploring usefulness and usability in the evaluation of open access digital libraries. *Information Processing and Management* 44(3) 1234-1250.

Vakkari, P., and Hakala, P. (2000) Changes in relevance criteria and problem stages in task performance. Journal of Documentation. 56(5), p540-562.

Wathen, C.N., & Burkell, J. (2002). Believe it or not: factors influencing credibility on the web. *Journal of the American Society for Information Science and Technology* 53(2) 134-144.

Wang, R.Y., Reddy, M.P. and Kon, H.B.(1995). Toward quality data: an attribute-based approach. *Decision Support Systems* 13(3-4) 349-372.

Webster, R. and Williams, P. (2005). An evaluation of the NHS direct online health information e-mail enquiry service: quality of health information on the internet. *Aslib Proceedings: New Information Perspectives* 57(1) 48-62.

Xie,H. (2006). Evaluation of digital libraries: criteria and problems from users’ perspectives. *Library and Information Science Research* 28(3) 433-452.

Xie, H. (2008). Users’ evaluation of digitial libraries (DLs): their uses, their criteria, and their assessment. *Information Processing and Management* 44(3) 1346-1373.

Yang, Z., Cai,S., Zhou,Z. and Zhou, N.(2005). Development and validation of an instrument to measure user perceived service quality of information presenting web portals. *Information and Management* 42(4) 575-589.

Yu, H. and Kaufman, D. (2007). A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. In: *Proceedings of the Pacific Symposium on Biocomputing* (World Scientific, Maui, 2007) 328-339.

Yushiana, M. and Rani, W.A. (2007). Heuristic evaluation of interface usability for a web-based OPAC. *Library Hi Tech* 25(4) 538-549.