Title: Using the repertory grid and laddering technique to determine the user's evaluative model of search engines.

S.E. Crudge, F. C. Johnson
Manchester Metropolitan University

Purpose

This study explores a method for the determination of users' representations of search engines, formed during their interaction with these systems. The purpose is to determine the extent to which these elicited 'mental models' indicate the system aspects of importance to the user and from this their evaluative view of these tools.

Methodology/Approach

The repertory grid technique is used to elicit a set of constructs that define facets within the mental model of an individual. A related technique of laddering then considers each of the user's constructs to determine the reasons for its importance within the user's mental model.

Findings

The model derived from the qualitative data comprises three hierarchical strata and conveys the interrelations between basic system description, evaluative description, and the key evaluations of ease, efficiency, effort and effectiveness. Two additional layers relating to the perceived process and the experience of emotion are also discussed.

Research limitations/implications

Ten participants is considered to be optimum for obtaining constructs in a repertory grid, but limits the findings to the context of the user group and the systems used in this study.

Originality/value of the paper

The methodology has not previously been used to determine mental models of search engines and from these to understand users' evaluative view of systems. The resulting model of key evaluations with the conjunctions of procedural elements suggests a framework for further research to evaluate search engines from the user perspective.

Keywords: Mental Models; Search Engines; Information Retrieval System Evaluation. Research Paper

Introduction

Fundamental to the progression of Web retrieval system development is the need to understand the way the tools are perceived by the end-users. Unlike the target audience of systems such as DIALOG, web searchers form a large body of 'ordinary' users, with little or no formal information retrieval (IR) training. Although studies such as Spink et al (2001) indicate that users' interactions with search engines are often brief, users will nevertheless form an opinion of the tools they have used, and the opinions will inform any subsequent choice of search engine. These impressions are unlikely to be formed in the abstract and may be underpinned by the users' 'mental models', the term often used for the user's internal representation of the system resulting from their interaction. This paper explores the value of these mental models in determining the criteria on which to base the evaluation of interactive retrieval systems from a user perspective. The rationale for this line of enquiry is provided in the following sections which define mental models, review methods for their derivation and identify requirements for evaluation criteria based on the user perspective. The remaining sections describe our approach to eliciting and the subsequent derivation of mental models, and the paper concludes with an analysis of the system descriptions, key evaluations and procedural elements that form the users' evaluative view of these systems.

Definitions of mental models

The definition of the term 'mental model' varies across the literature, and has been the subject of much debate in the field of human computer interaction (Staggers & Norcio, 1993). In an information system context, the term most frequently refers to the conceptualisations of systems formed by users to assist their understanding of how the system works. Norman (1983) clarifies the distinction in the terminology between the user's mental model of the system and the conceptual model of the system designer's view of the target system. It is expected that conceptual and mental models must be similar in order for the interaction to prove successful, and Staggers and Norcio (1993) extend the purpose of the conceptual model to facilitate the development of a user's congruent corresponding mental model.

Previous studies have often concentrated on the inaccuracies within mental models. Brandt and Uden (2003), for example, found that users expect semantic meaning to be derived from web sites by the engines. Furthermore, the study found that the difference between directory and search is not fully understood, there is little perseverance for scanning of result lists, and little understanding of the search index overlap with the index of other engines. Others however claim that mental models should be examined for the expectations they set, rather than for their accuracy (Seadle, 2003). The construction of a plausible mental model is thought to be the source and to govern users' expectations about the effects of actions and can guide the way the system is used and how feedback is interpreted (Van Der Veer, 1989). If this is so, the quality of a user's interaction with the system will depend not only upon the user's mental model being congruent to the conceptual model but more importantly upon the functionality of that model. Understanding human interactive behaviour through these models would seem to be a promising approach to inform the design of effective interfaces. Yet, despite the interest, there is a lack of a strong theoretical basis for mental models (Fetzer, 1993; Johnson-Laird & Byrne, 1991) and research with respect to mental models, system and interface design is not yet well developed.

Eliciting mental models

The difficulty in conducting research based on the notion of mental models stems largely from the lack of a standard technique for derivation of the models. Many popular techniques involve generalisations, such as requiring a user to draw a picture representing the technology, or to describe it in a few sentences. This approach to its determination stems from the belief that users' mental models are formed on analogies to similar technologies (Staggers & Norcio, 1993). A study by Slone (2002) asked participants to explain how the Internet and online catalogues worked. The resulting online catalogue models were often obtained through comparison with other system types, whereas the Internet was often given 'magical' or human characteristics, felt by the researcher to be suggestive of fragmented or immature models. Ratzan (2000) reported a study of 350 participants who were surveyed to determine views of the Internet. Metaphors were used frequently, and a view of the Internet as a disorganised library was common, but expert users suggested various and sometimes metaphysical metaphors such as 'fractal' and 'new dimension'.

Although it is reasonable to assume that analogy with other technology will assist mental model development, it is more likely that the precise detail of a model will be formed as a result of interaction with the systems (Norman, 1993), and that cause and effect chains are an integral part of this (DeKleer and Brown, 1983). This being the case, much information regarding mental models could be obtained from gathering usage data, such as transaction logs. Transaction log analyses have been undertaken for several search engines (Spink et al, 2000; Silverstein et al, 1999), and such studies indicate low levels of query reformulation, and a reluctance to view beyond the first few pages of retrieved items.

Few studies make an explicit link between transaction logs and mental models. Moukdad and Large (2001) described user mental models of the WebCrawler search engine through examination of a sample of the queries posed to the engine. The study speculates that users pose questions to the engine because they view it as they would a human respondent. Muramatsu and Pratt (2001) investigated models by determining the users' understanding of the system interpretation of queries. The results indicated that the participants expected engines to combine search terms with ‘OR’ rather than ‘AND’, and expected term suffix expansion. Little knowledge of stopwords was exhibited, and only slightly more understanding of term order variations was detected. The authors concluded that the participants’ models were naïve and incorrect. Such findings appear to indicate that users’ models can be incongruent with that of the system and impede on successful interaction and use.

Users’ evaluative mental models

The extent to which mental models can provide a better understanding of the user’s evaluative view of a retrieval system forms the focus for this study. The premise of a mental model is that users will form a mental representation of the system, however brief the actual interaction has been. Furthermore, as this model is thought to be derived during interaction and will subsequently facilitate successful system use, those system aspects included in the model may form the basis of the users’ evaluative view. In other words, the user’s evaluative view will be formed on the aspects and features that facilitate the accomplishment of a search task through interaction with the system.

Our investigation of the system aspects of the user’s mental model (a system model) has a twofold value. System designers will be more aware of the aspects users consider when making their initial, but critical, assessment of the system. Furthermore, the user model of the system can be used to identify meaningful user-based criteria, for use in the evaluation of interactive retrieval systems. Evaluation of these systems requires assessment of performance, not only in terms of retrieved output but also in the support the system provides to facilitate the users’ search process. Often the user’s assessment of system ‘success’ is obtained from some combination of ratings along criteria such as system effectiveness, efficiency, interaction/usability and overall satisfaction and/or success (for example, Johnson 2001; Su, 2003). However, Xie (2003) highlights the complexity of criteria such as 'ease of use', and the assessment of such a dimension poses problems when no standard or commonly agreed operationalisation exists. To address this, in an evaluation of online systems, Xie first asked the participants to define the concepts of ease of use and user control, before recording perceptions of the concepts for each system feature used during the process of an online search. A feasibility study (Johnson et al, 2003), conducted to identify a framework to evaluate search engines support for the search task, similarly identified the need to better understand the user's evaluative view when defining assessment criteria for system features and impacts.

Eliciting evaluative mental models

The method chosen for the determination of the users' mental models was the repertory grid approach in conjunction with laddering technique. Grid technique originates in the field of clinical psychology (Kelly, 1991), and involves the elicitation of the system of interrelated constructs, thought to represent the hypotheses derived from human experiences, which govern expectations of the world. Constructs are defined as "a way in which some things are construed as being alike and yet different from others" (p. 74). Fransella, Bell and Bannister's (2003) manual on the design and usage of grids suggests its popularity as a method for eliciting personal constructs. Although relatively few studies have employed grid technique in the field of IR, the method has been used to model information space (McKnight, 2000), to determine mental models of IR (Zhang & Chignell, 2001), and in the classification of text types (Dillon, 1994; Dillon & McKnight, 1990) and digitised photographs (Burke, 2001). It is used in this study to determine the users' evaluative view of search engines, allowing the participants to state as many or as few system aspects as they wish, without researcher influence. The laddering technique explores the reasons why a given system aspect is important and results in a full interpretation of each concept.

Kelly believed that construct systems are hierarchically organised and interrelated by cause and effect, with some constructs being central to the beliefs of an individual. These core concepts can be visualised as forming the topmost points of a pyramid, with the lower positions filled by the system of interrelated constructs. During laddering, the interviewer starts at any point within this system, termed the ‘seed item’, and using a series of probing questions the participant is guided up, down and across the (hierarchical) construct system (Rugg et al., 1999). This method is essentially a combination of Hinkle’s (1965) laddering technique used to move upwards within the hierarchy, with Landfield’s (1971) pyramid technique used to move downwards in the hierarchy. It has now become standard for the term ‘laddering’ to refer to the combined method.

Mental models are of course highly individual, but for the given technology of Internet search engines, within a specified contextual definition, it is reasonable to assume that a model could be determined from which all individual models would be taken. This model will be termed the complete mental model, and is defined as the smallest set containing all individual model subsets. It must be reiterated that, as for any study involving real users, the determined model will be subject to the contextual implications of the user group and technology from which it derives.

Research questions

The research questions posed are

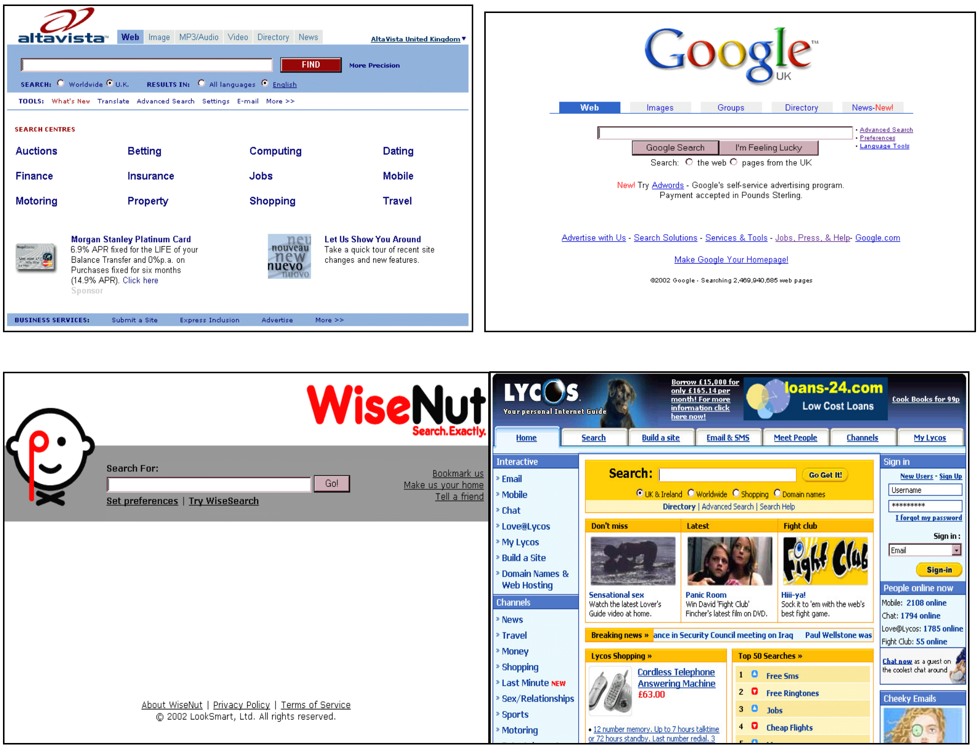
1. *Can a suitable evaluative representation for a user mental model be obtained using laddering technique, based on a set of constructs determined from a repertory grid.*

2. *Can the resulting mental model inform system design and further contribute to the evaluation criteria for use in assessing search engines.*

Method

Ten first year undergraduates were recruited for the study during the second week of their first year. A small sample size of between six and ten is commonly used when implementing a repertory grid investigation (Dillon & McKnight, 1990; Hassenzhal & Trautmann, 2001) and Moynihan (1996) and Dunn (1986) using samples of fourteen and seventeen respectively found that no new constructs were elicited after the tenth participant. In this study data were collected from ten participants from the Department of Information and Communications at Manchester Metropolitan University. Data were collected on an individual basis, and the sessions involved three stages: introduction to a selection of search engines during a familiarisation session, a tape-recorded interview during which constructs were generated for inclusion in a ratings grid, and exploration of the constructs using probing questions.

Familiarisation session. To identify the engines for use, a number were profiled to determine a small set representative of common search technologies at the time of the study. The engines chosen were AltaVista UK, Google UK, Lycos UK, and Wisenut, (see Figure 1) and these formed a set of ‘elements’ from which to elicit the constructs. Each participant searched using each engine for information to satisfy a chosen coursework assignment, thus ensuring motivation and realism of task. Time spent with each system was constant across the set, but the order of presentation of the systems varied across participants, to reduce learning effects.



6.9% APR fixed for the LIFE of your Balance Transfer and 0%p.a. on Purchases fixed for six months (14.9% APR). Click here

Take a quick tour of recent site changes and new features.

New! Try Adwords - Google's self-service advertising program. Payment accepted in Pounds Sterling.

©2002 Google - Searching 2,469,940,685 web pages

© 2002 LookSmart, Ltd. All rights reserved.

Borrow £15,000 for only £165.14 per month! For more information click here now!

Watch the latest Lover's Guide video at home.

Win David 'Fight Club' Fincher's latest film on DVD.

Sock it to 'em with the web's best fight game.

ance in Security Council meeting on Iraq Paul Wellstone was

• 12 number memory. Up to 7 hours talktime or 72 hours standby. Last number redial. 3

Chat now as a guest on the coolest chat around

Figure 1: Screenshots of the search engines

Construct elicitation and grid completion. Participants gave an overall rating of success for each search engine, taken immediately after familiarisation. The method of dyadic elicitation was then used to generate constructs for use in the qualitative study. During this process, participants considered the search engines in pairs, and stated either a similarity or a difference between the members of each pair. The opposite of the stated similarity or difference was then obtained to form a construct, represented by a five-point scale along which all engines were rated. During elicitation, an additional engine, the participants' perceived 'ideal' search engine, was introduced; 'ideal' elements are commonly included in grid studies where element number is low (Whyte & Bytheway, 1996; Hunter, 1997). Pairs of engines were presented until no new constructs were elicited, and participants were then allowed to select further pairs of elements and provide additional constructs if they wished.

Laddering. The grid completion phase provided quantitative data and a great deal of qualitative data relating to more detailed exploration of the constructs was also obtained using laddering. Corbridge et al. (1994) emphasise that the probes used during the process should be standardised. The general rules given by Stewart and Stewart (1981) recommend use of the probe 'why is that important to you?' to take the participants higher up their pyramids, while probes such as 'how is it different?' will move lower. A common strategy begins with the elicitation of a construct using a triad. The participant is then asked to identify which pole of that construct they prefer, and then to state why they prefer it. This can be repeated, and the participant continues to move higher up the construct pyramid. Once the participant is no longer able to move upwards, the interviewer can return to the original construct and begin a series of probes that will assist the move downwards. One such method requires the participant to state how the two poles of the construct are different from each other. An outline example demonstrating laddering of the construct 'Interface Simple / Cluttered' is given in Figure 2. The type of probe being used at each stage is indicated by the italicised comments, and a visual representation of the construct hierarchy is also provided beneath. This type of diagram was used during data collection for this study, to record the basic laddering information in note form.

Laddering Upwards

Interviewer: Considering the construct of interface simple / interface cluttered, which would you prefer? *[Determining the positive pole of the construct]*
Participant: A simple interface
Interviewer: Why would you prefer a simple interface? *[Probe to move upwards]*
Participant: Because it is easier for me to see where I have to type in the words
Interviewer: Why is that better for you? *[Probe to move upwards]*
Participant: Because then it's quicker to search.

Laddering Downwards

Interviewer: Thinking about the difference you just mentioned of a simple or cluttered interface. Can you think of any ways in which simple and cluttered interfaces are different? *[Probe to move downwards]*
Participant: A cluttered interface has lots of writing on it.
Interviewer: Can you explain what you mean by writing? *[Clarifying answer]*
Participant: Links to other things
Interviewer: Can you think of any other ways in which simple and cluttered interfaces differ? *[Probe to move sideways]*
Participant: Cluttered interfaces have lots of adverts.

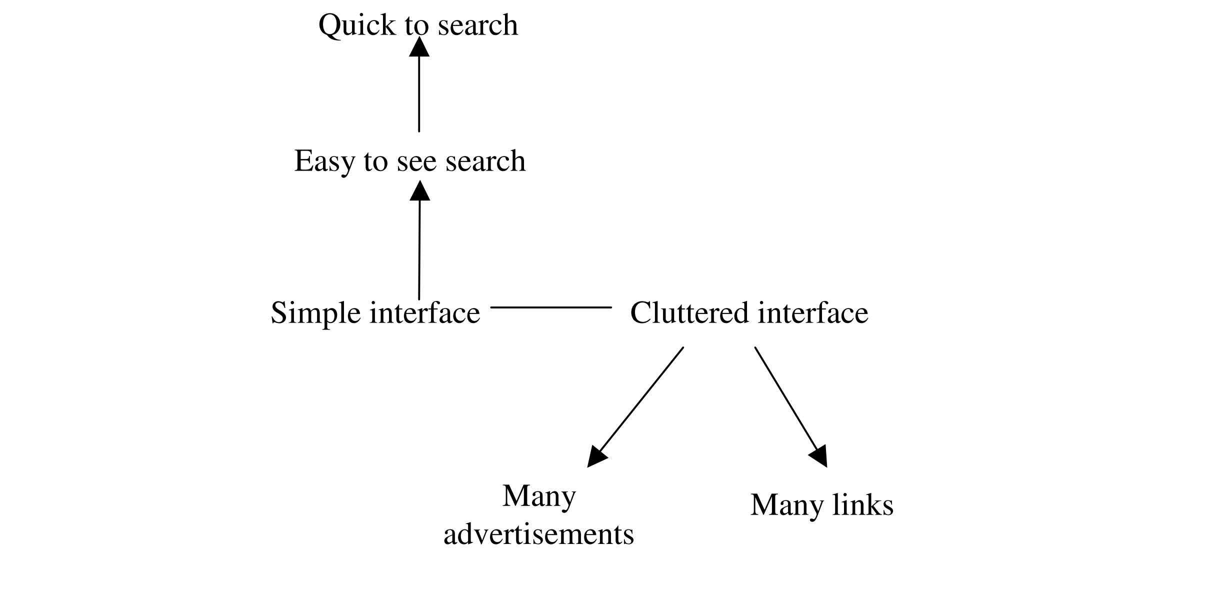


Figure 2: Elicitation of ladders with construct 'Interface simple/ Interface cluttered'.

Data Analysis

Detailed analysis of the suitability of the repertory grid technique for eliciting a mental model of search engines was presented in Crudge and Johnson (2004). The validity of the technique in the given context was established through the use of the method to generate a finite set of constructs that discriminated tolerably well over the set of search engines, and with the majority of constructs clustering closely with the overall success rating. The grid data formed only a small portion of the data obtained from the study, with the main analysis taking a qualitative approach and being based on the more substantial laddering data.

Following transcription of the tape-recorded interviews, the raw construct set was used to provide a partial template to facilitate first level coding. The data was divided into 479 short segments, indexed by 65 different codes. Atlas/ti (Muhr, 1997) was used to enable grouping of the coded sections into themed subsections and the hierarchical consequential relations between data segments were identified using the probing questions of the laddering technique. This stage was derived primarily from the means-end chain analysis method proposed by Reynolds and Gutman (1988) for the analysis of laddering data, but also corresponded to the axial coding phase of Grounded Theory. An example of a consequence chain derivation is provided by Figure 3. The direction of the arrows indicates the direction of the implication, with the left hand side corresponding to the lowest levels of the hierarchy, and the probe 'Why is that important to you?' being used to move across to the higher levels of the hierarchy on the right.

Participant: This is better, it's at the top [refers to directory link].
Interviewer: **Why** is it better at the top?
Participant: More aware of it at the top. When you see the page you see that first.
Interviewer: **Why** do you want to be more aware of it?
Participant: So I don't waste time looking to see… I want to know it knows my query.

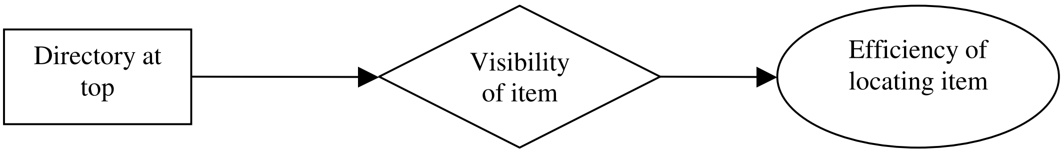


Figure 3: Derivation of a consequence chain.

All the consequence chains were examined, together with the themed groupings already identified, and a generalised consequence chain was determined. This represented all the possible hierarchical interrelations between the main data types, with a large proportion of the data appropriately assigned to one of three categories, ranging from the lower hierarchical levels of *basic description*, through the middle levels of *evaluative description*, to the highest levels termed *key evaluations*. The generalised chain is included as Figure 4, with the causal relations indicated by the arrows, the thickest of these providing the main pathway through the hierarchy. The thinner solid lines indicate the possibility that statements from one data type could cause statements drawn from the same data type. Finally, the broken lines indicate the presence of affective statements within consequence chains. There was a substantial portion of data pertaining to emotional responses to the systems observed to occur at a variety of points in the hierarchy.

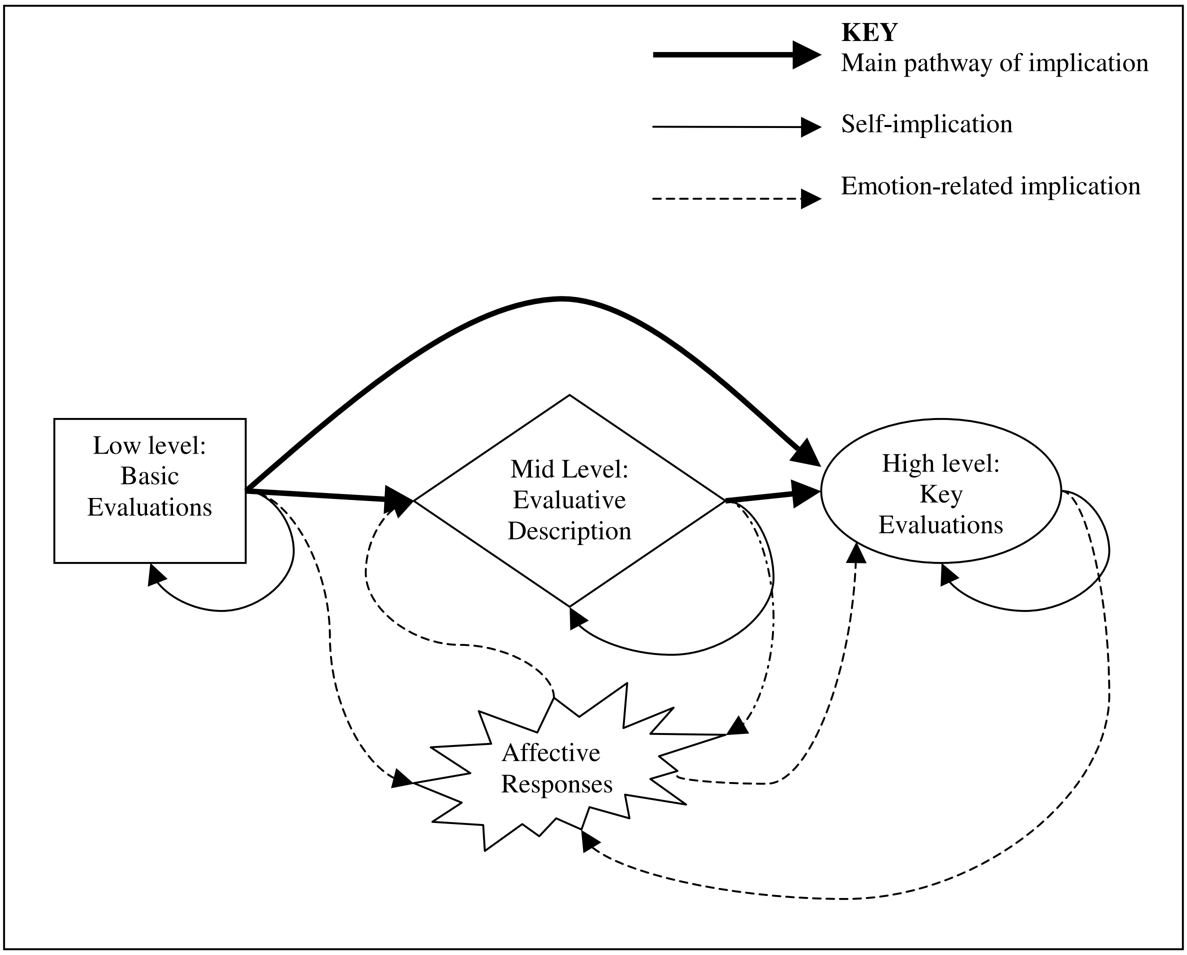

KEY
Main pathway of implication
Self-implication
Emotion-related implication
Low level: Basic Evaluations
Mid Level: Evaluative Description
High level: Key Evaluations
Affective Responses

Figure 4: Generalised consequence chain.

Thematic Discussion

The final code types identified serve to divide the data into three main hierarchical areas comprising basic description, evaluative description, and key evaluations.

Basic description

The basic description forms the lowest levels of the consequence chains derived from the participants during laddering. The data from this section is characterised by its focus on description, and divides into description of the screens and the features of the systems. Table I provides the number of participants reporting aspects of screen layout or features.

Aspect	No. of Participants
Screens	**10**
Front page	8
Result pages	9
Features	**10**
Named features	9
General presentation of features	9

Table I: Participants providing data of basic description.

Screens. Three screens are described, the main entry page, commonly called the front page or interface, the result pages, and the advanced search page. The main issue emerging for the design of the front page related to the style, which usually reflected the streamlined or portal appearance. Participants gave description such as ‘busy’, ‘plain’, or ‘cluttered’, and discussed the presence of links, writing, adverts and ‘stuff’ on the page. Plain front pages were usually preferred, but one participant felt such pages could have too few colours. The issue of colour resulted in difference of opinion, with other participants preferring fewer colours. The colour was also referred to as mellow, garish or heavy. Five participants referred to the search box, which suggests that this featured strongly in the mental models of these participants.

All but one participant discussed the layout of the results page. In a parallel with the front page, participants distinguished between plain and busy pages, with only one preferring the busier variety. Definition of plain and busy varied, relating to font type and size, inclusion of URL and other information, use of numbering, and size of site descriptions. Other result page aspects included the need for a statement of the results quantity, use of colour to identify visited URLs, and a link to further pages of results.

Advertising. Advertising was discussed by eight participants, with varying reference to location on front or result pages. The participant set varied in attitude to advertising; although several participants felt that adverts should not be present there was some degree of acceptance, and one participant even felt advertising could be a positive issue if it took the form of a joke or cartoon. The quantity of advertisements was connected to the overall style of the interface. When considering advertising on the result pages, the two main issues were location and relevance. Positioning immediately prior to search results was a bad aspect; better presentation had advertisements grouped together and at the side of the page. While participants mentioned that relevant advertising might be acceptable, there was some disagreement over the definition of this, with one participant stating that a link to a store selling books on the search topic was relevant, and a second participant giving the same example as not relevant. Pop-ups and moving advertising were not favoured, and the colour, size, and ‘subliminal’ nature of the advertising were also mentioned.

Functionality. All participants discussed features, with specific features mentioned including conceptual clusters, sneak-a-peek, directory, advanced search, image search, language facility, news, and e-mail. Several general issues pertaining to functionality were raised, including the quantity and variety of options available, and the relevance of features to searching. Only one participant stated a preference for the inclusion of non-search related options. The presentation of the features was also referred to, and the tab-style was usually favoured, but one participant preferred a drop-down menu. More commonly, participants referred to the location of the features, or the location of the point of access to the features. Location at the top of the page or on the front page was often preferred.

Evaluative description

Table II provides the number of participants reporting each of the three main areas of evaluative description, namely readability of the screens, visibility of items including the search box, and the content of the results. These are areas that typically appear lower down the consequence chains, but are not purely descriptive.

Area of Evaluative Description	No. of Participants
Content criteria for results	**10**
Quantity	9
Relevance	9
Precision/ ranking	7
Readability	**7**
Front page	4
Result pages	4
Features	3
Visibility	**7**
Search box	5

Table II: Participants presenting evaluative description.

Readability of the screens was most affected by the choice of a streamlined or portal style, and the use of colour; plainer interfaces were more readable, while heavy or large amounts of colour were ‘hard on the eyes’. Readability of features such as the tabs or pull-down menus was also discussed, with size and colour affecting this. Visibility was an issue, especially for the search box and access to features such as the directory. The location most affected the visibility of access to features, whilst the interface style was commonly stated to affect search box visibility. Locating access to features at the top of the page or on the front page would increase visibility, whilst the adverts, writing and 'stuff' on an interface caused reduced visibility, especially noted for the search box. Several participants discussed the use of colour to highlight terms or the statement of number of 'hits', thereby rendering them more visible.

The content of the results was commonly discussed, but the criteria upon which results were assessed varied greatly. The table indicates areas where a degree of commonality occurred across the participant set, but a number of more individualised criteria were also specified. Some of the content criteria seemed heavily dependent on each other, and were not clearly delineated within the participant's mental models. There is interlinking of concepts such as quantity, precision, ranking, and relevance, which is also complicated by the inclusion of the process element of refining. A selection of participant comments illustrating the interlinking of these issues is provided as Figure 5. In addition to these main issues, raised by a high proportion of the participant set, smaller numbers of participants also raised a variety of other issues, including presence of familiar results, utility, and quality of retrieved sites. Only one participant discussed recall.

- "Because that's why you're visiting the search engine in the first…. You want relevant useful results, or at least results that are going to make you think about searching on a different term, or that are heading the right way towards finding the answer that you want."
- "I wouldn't mind how many results there were so long as they were all relevant to what I was looking for. But if they were totally unrelated or they were, they weren't what I was looking for, then obviously the less results the better really, less but pertinent results."
- "If I was left with, say, 12 results, I would expect that to be more in depth and detailed and more useful for what I was looking for."
- "You don't want to have, well the ideal thing is to possibly have fifty results or something, you don't want any more than that otherwise you'd be…. So of course if you

don't have it, you don't want to have 300,000 results or something and if you've got no way of reducing them you're just going to be lost."

Figure 5: Selection of comments relating to relevance, refinement & quantity of results.

Key evaluations

Laddering of data provided within the lower categories of basic or evaluative description often resulted in a 'key evaluation'. Laddering procedure guides participants from their peripheral beliefs towards their core beliefs, with data increasing in abstraction from the context. For this study, the core beliefs are termed 'key evaluations' and were identified as the highest levels within the consequence chains, and often terminated the chains.

Key evaluation	Definition	Number of Participants
Effectiveness	Any combination of 'content criteria'	10
Efficiency	Time taken both overall or at stages	10
Ease of use	The ease of achieving aims	9
Effort	Physical 'amount' required	5

Table III: Participants reporting of key evaluations.

The terminology chosen for the four key evaluations reflects the ideas of literature and research, and although the terms efficiency and effort can have a more complex interpretation, for the purposes of this study they simply represent participant statements such as 'time taken' or the 'amount' a task must be performed. Effectiveness is taken to be the often highly individualised combination of the content criteria. The intended interpretation for each key evaluation is provided by Table III, together with the number of participants reporting the concept. Ease and efficiency occur quite frequently in the data, with all participants referring to efficiency and nine referring to ease. Effort occurred less often, with only half the participant set referring to this. Some examples of user statements relating to the key evaluations are provided in Figure 6. Identification of several co-occurrences of key evaluations within the data suggests that the concepts may be interlinked. However, there is inconsistency in reporting that makes it impossible to draw conclusions about a possible hierarchical order for the concepts. For the purposes of discussion, references in the data have been explicitly separated as far as possible.

Ease

- "Oh, I just found it user-friendly, I just found it nice. Because sometimes I must admit, I can just close a window and close the whole damn thing, you know, and I've got to go back again, whereas with that it's easier not to do that, isn't it." [sneak-a-peek]
- "I think AltaVista is near to my ideal engine, it's very good, easy to find the result. The way you search is very good."

Efficiency

- "When you're looking for something you don't want to spend hours and hours searching for it, you just want to find it and get on with what you're doing basically."
- "What I did like was on the Wisenut one you could take a preview of the actual site…if you're looking for something quickly, saves you having to like click forwards and backwards and that sort of stuff."

Effort

- "You have to work out a little bit more yourself more words to put in."

Effectiveness

- "I expect it to find relevant data, I expect it to find all the data, because it's supposed to be powerful, and I expect it not to give rubbish providing your search command is reasonably precise."

Figure 6: Selection of comments relating to key evaluations.

Ease of use. Several categories of ease of use were described, namely the ease of use of search features, ease as increased by search features, and the ease as affected by the design of the screens. When discussing the ease of use of search functionality, one participant related the complexity of the advanced search to ease of use, and felt that a complicated advanced search would result in non-use. Wisenut's sneak-a-peek feature was stated to be easy to use, either navigationally or by reducing errors. Term suggestion features were also easy to use navigationally. Comments such as “you just click” were common explanations for ease of using conceptual clusters. The location of conceptual clusters at the top of the page increased their visibility and thus affected their ease of use. Finally, one participant discussed that the style of presentation of features would have an effect on the ease of use, preferring drop down menus to tabs, with the vertical list approach of the pull-down menu being more readable and so easier to use.

‘Ease as increased by search features’ was referred to by three participants as ease of use of the general search mechanism itself. Screen design issues such as quantity of information, colour, and advertisements impinged on the general ease of use, and one participant elaborated that for an interface with many links, it became more difficult to pick things out, and so was harder to use.

Efficiency. Efficiency was commonly reported during laddering, with all participants stating the time required to be a consequence of at least one lower level descriptive element. A long time taken whilst using a search engine was always a negative aspect.

Eight participants gave time saving as a reason why the results content was important. The relevance, quantity and precision of the results were all stated to have an impact on the time required, and one participant mentioned scrolling through the results as the reason why the time was increased. Another participant expressed a dislike of the inclusion of PDF file types in the results content, and explained that these could take too long to load in.

The layout of the result pages was also stated to lead to extra time being required. Colours to indicate visited links would save time by reducing unnecessary revisiting of sites. A greater number of lines in the site descriptions would speed up the assessment process and the navigational aspect of clicking on titles to visit a site was also felt to be quick.

The evaluative descriptions of readability and visibility both affected the time required when using the front page. One participant who found the 'busy' interfaces harder to read felt that this impacted on time. The extra time required to locate the search box if it was surrounded by other information, was also highlighted. Advertising slowed down two participants, who cited pop-ups and moving adverts as causes of this problem. The visibility of the adverts was another cause of time expenditure, for pop-ups this was explained as the time required to close them down.

Features were often stated as time-saving, with the reasons usually linked to the perceived use of the feature. Two participants stated that the cache feature saved time by ensuring access to sites even when 'down'. Half the participant set felt that sneak-a-peek would save them time; the reduced need to open a new page, reduction of navigational forward and backward clicking, and the use for relevance assessment were reasons provided. Categorised results/ term suggestions also saved time, either by allowing quick access to a subset or to quickly obtain more relevant information.

The presentation of the features was also linked to the time required. The 'quick launch' access to news at AltaVista, use of tabs or clicking to access things, and the location of functionality would all save time. The location of the directory was mentioned by three participants; placement at the top of the page increased visibility and reduced time. Finally, the location of pull-down menus or tabs at both top and bottom of page would reduce the need to scroll, thus saving time, and stated by one participant.

Effort. Effort was the most difficult key evaluation to identify from the data, and is evidenced by statements such as ‘amount required’. The effort is more easily understood as a conjunction with process stages and actions, for example the amount of scrolling required. Participants variously discuss the amount of formulation required, the amount of navigation, the scrolling as linked to location of features and results precision, and the changing of pages, especially during relevance assessment, as linked to features that reduce it. Finally, the amount of refinement required as related to the content of the results was also discussed.

Effectiveness. Many participants were observed to combine various content criteria in an individual manner in order to define a ‘good’ result. The content criteria for effectiveness are more complicated than just having the information you want. There are simple methods the users employ to judge the relevance of a site at a glance. For example, several participants judge sites to be ‘right’ if they are the same as those retrieved from other sites. Others use term proximity as observed in site descriptions for relevance assessment. The presence of such shortcuts to relevance assessment perhaps stems from the users in the study providing general constructs relating to the overall effectiveness, rather than criteria for an individual site’s relevance. However, the main issues for effectiveness were the quantity and relevance of the results.

The issue of results’ quantity was closely linked to the precision, relevance and ability to refine. Participants were often unable to separate these concepts out in their discussion. Several participants equate the quantity of results with irrelevance, expecting a greater number of irrelevant results to be present in a larger retrieved set. To this extent, the quantity of results influences the perception of them by participants and the attitude to refining then becomes important, with some participants being more prepared than others to formulate or refine. The interrelations between these issues, and the combination of the criteria to produce an overall effective result, are complex and highly individualised, as would be expected.

The Mental Model

A summary model of this data, as presented in Figure 7, provides a representation of the users' mental model of the systems comprising basic description, evaluative description, and key evaluations. The analysis of the majority of the data set identified these hierarchical delineations within the data, related as consequence chains elicited during the laddering process. During analysis, it became apparent that there also existed a prevalence of emotional and procedural statements, which were not directly requested from the participants. These were analysed and are depicted in Figure 7 as layers overlapping and interacting with the central mental model taking the traditional pyramid form with key evaluations defined by the attributes and values lower in the hierarchy. The insight the mental model provides into the user's evaluative view of the system is discussed in these final and concluding sections.

"Take in Figure (No.7)" **Figure 7: The users' hierarchical mental model**

The hierarchical evaluation layer

A pyramid is a common visual representation for laddered data, and in traditional terminology, the lowest levels are termed 'attributes', the middle levels are 'consequences', and the highest levels are 'values'. There are usually more attributes than consequences, and more consequences than values, hence the visual use of a pyramid representation. This is also the case for the data set here, with the key evaluations providing the core concepts, positioned at the top of the pyramid and of fundamental importance to the participants. The criteria of the key evaluations of ease, efficiency, effort and effectiveness are used often in the literature on IR system evaluation. What is of interest in this study is their derivation as responses to notable system features and aspects. It is not surprising that, in this respect, users are largely unconcerned with the inner workings of the engine, its index and search algorithms. Detailed knowledge of these would not be expected from end users. In the same respect nor is it surprising that the key evaluations are perceived through the more obvious aspects of search results, screen design and functionality.

In the context of this model it is evident that users do hold an evaluative view of the search engines. The criteria on which this view is based are influenced by fairly complex interrelated reactions to aspects and features of the systems, and not solely on the system output of retrieved items. The laddering method has enabled the identification of several key evaluations, namely effectiveness, efficiency, ease and effort, with strong influencing factors on these expressions arising from the content of retrieved results, the presentational aspects of screen design, and the functionality of features. The causal impact of presentational aspects such as colour, adverts and location on user statements related to ease and effort of the search may be of interest to system design.

The affective layer

Participants had not been required to make their process or emotions explicit at any point during the interview procedure, but nearly every participant did so. This affective stratum is shown in Figure 7 containing the five types of emotional statements elicited from the participants during the interview process, namely distraction, confusion, frustration, boredom and overload. These emotional responses were identified as stemming from a variety of lower level data, and were usually the final members of a consequence chain. However, distraction also appeared as a cause of other aspects. Examination of the main causes of distraction identified a high proportion of presentational aspects, such as adverts, colour and amount of writing. Confusion similarly resulted from presentation aspects, such as the fonts and formatting, readability and clutter on interfaces. Advertising and visibility of features were the main causes of frustration, and the time taken was also a strong influence on an expression of frustration. Boredom related to having to perform a task repeatedly, such as refining or reading through imprecise results. The exploitation of the web-based medium by use of colour and layout was stated to increase interest. The causes of information overload were of two varieties, a profusion of retrieved sites, or an abundance of information provided by the search service itself.

The influence of presentational system attributes on a user's emotional experience is again expected to be of interest to system designers. It is quite likely that most users may cease their interactions with search engines with only emotional memories, and without themselves fully understanding the causes of the emotions at a conscious level. It is not suitable to suggest that the users' evaluative view is based strongly on emotional responses; psychological theory of emotion is required to inform the interpretation of the experience of emotion, and this is beyond the scope of this paper.

The procedural layer

The procedural layer in Figure 7 contains the data pertaining to process phases and actions, as derived from the participants' data. As described earlier, following coding, theming, and examination of consequence chains, the data was re-coded to identify process elements. The procedural data presented within this section is therefore not part of the hierarchical data pyramid. Although it was an intention of the analysis to identify the procedural perceptions of the participants, it was not anticipated that any level of detail would be determined. It was therefore unexpected that when taking a cumulative view of the data set, a detailed set of process stages and action statements did emerge.

These process statements derived from the data were grouped into three main phases, namely query input, results and refining, in line with the traditional IR model outlined by Saracevic (1996). Further subdivisions are indicated in Table IV, together with the frequency of reporting and action statements found to be common across process phases. These actions involved location of items such as the search box or features, reading the screens, typing in, and the navigational actions of scrolling, activating items, and changing pages.

Process phases and stages	No. of Participants
Query input phase	**8**
Formulate the query	8
Perform an advanced search	4
Results phase	**10**
Manipulate the results	4
Visualise the query	2
View the sites	5
Assess the sites	7
Refining phase	**10**
Improve the relevance of the results	5
Reduce the quantity of the results	6
Action statements	
Read screen	9
Locate item	8
Type in	7
Navigational actions	8
Change page	7
Activate item	7
Scroll	3

Table IV: Participants referring to process phases and stages.

The process statements are examined here for the interpretation they provide for the users' key evaluations. This examination took the form of identifying the conjunctions of process elements with mention of specific features, indicating their perceived uses for search features. These might include, for example, use of conceptual clusters to ‘refine the search’, use of directory links to ‘visualise the query’ or use of advanced search to ‘reduce results quantity’. The second point at which procedural elements were identified within the data set was as qualifying statements in conjunction with key evaluations. For example, a participant might discuss that it was ‘easy to refine a search’, thereby providing a conjunction of the key evaluation of ‘ease’ with the process phase of ‘refinement’.

A full chart of possible conjunctions of process with key evaluations ease, efficiency and effort is provided as Table V. Combinations occurring in the data are indicated by a tick, whilst combinations marked with a cross did not occur at all in the data set. The frequency of identified combinations is also provided. The main divisions of query input, results and refinement phases receive a count of the number of participants referring to the category in general, and this may have been in addition to conjunctions with one or more specific subcategories, as indicated. The key evaluation of effectiveness, in the context of this study taken to be some combination of the content criteria, was not seen to occur in conjunction with the majority of the procedural elements. However, while participants did refer to effectiveness without the use of refinement, and effectiveness after refinement, this is not included in the table.

	Ease	Efficiency	Effort
TOTAL	9	10	5
Query input phase	✓ (6)	✓ (2)	✓
Formulate query	✓ (1)	✓ (1)	✓ (2)
Advanced search	✓ (2)	✗	✗
Results phase	✓	✓	✓
Manipulate results	✓ (1)	✓ (2)	✗
Visualise query	✗	✓ (1)	✗
View sites	✓ (2)	✓ (6)	✗
Assess sites	✓ (3)	✓ (5)	✓ (3)
Refining phase	✓ (3)	✓ (5)	✓ (2)
Reduce quantity	✓ (1)	✗	✓ (1)
Improve relevance	✓ (1)	✓ (3)	✗

Table V: Reports of key evaluations in conjunctions with process stages.

From the total possible set of 24 specific process/ key evaluation conjunctions, two thirds were observed in the data. The highest percentage was observed for ease of use, with seven of the eight possible conjunctions occurring in the data. For efficiency, six of the eight conjunctions were reported, but the number observed for the effort evaluation was noticeably lower, with only three conjunctions reported.

The emergence of key evaluation and process conjunctions suggests that user evaluation criteria of ease, efficiency and effort have shades of meaning when considered with the procedural statements. The structured elicitation of the mental model of the system links the system attributes in a causal relationship to users' key evaluations, and these are themselves qualified by consideration of the task process in which the user is engaged. Thus, when used in a system evaluation, it is expected that users' ratings on these key evaluations will not simply be abstract reactions to the system, but direct responses from an assessment of key aspects. Furthermore, assessments on these criteria need to be taken at or with consideration for the identifiable process stages. A system may receive an overall rating for ease of use, but obtain very different ratings for ease of use when considered at the process stages. The relationship in the model between evaluation, system description and search process, leads us to expect that user evaluation criteria based on the evaluation/process conjunctions could provide a meaningful user assessment of the system. In the final section the further research necessary to develop and utilise these criteria is presented.

Conclusions and further research

The model here is taken to be the compilation of the individual models of ten participants, and it is expected that this model is complete, in the sense that the addition of any further participants in a repertory grid study would not generate any new constructs. Furthermore, previous research suggests that mental models in general will increase in accuracy and completeness as the experience level of an individual increases. The sample had moderate levels of experience, with nine out of ten stating average or above average search engine experience, and eight out of ten stating average or above average Internet experience. It is thus expected that the participants in this study have presented reasonably complete models, with reasonable accuracy.

The value of the model determined here is in essence its representation of the users' evaluative view of the search engines. The elicitation method made use of similarity and differences between systems to draw out comments defining the users' perceptions of the tools. As the derivation was based on comparisons, so the resulting model takes an evaluative context, and each individual interview essentially resulted in an evaluation of the system by that user. This study has not been concerned with the actual ratings or assessments given to individual engines, rather its interest lies in the resulting explanatory model. The hierarchical mental model derived from the laddering data explains the importance of users' descriptions of system features by the key evaluations. This provides us with more understanding of the system features that are important to the user and why. Such understanding is of interest to system designers, with a view to developing a system to which users will make that all-important initial positive judgement.

Viewed another way, our model makes explicit the many potential determinants, or at least influences, on user evaluative criteria, such as ease of use, which hold value for further system evaluation studies from the user perspective. In particular, the conjunctions of procedure with key evaluations suggest the value of developing a methodology for user based system evaluation, using a set of criteria such as ‘ease’, ‘time’ or ‘effort’ taken at identifiable process stages. This aspect of the findings, whilst promising for the development of evaluation criteria, requires further investigation in a continuation of this study beyond the ten participants. Further research is needed to ascertain the extent to which the evaluative view is held in a larger sample and to begin to verify our suggestion that key evaluations will hold varying degrees of importance depending on the process stage of the user. Further research using the repertory grid may also usefully explore whether new constructs arise from variant user groups and/or whether variation in the element set would allow for the emergence of new meaning in the constructs.

Seeking an approach to evaluate the retrieval system objective in supporting the search, its interactivity and functionality, is highly complex. A typical usability study may be able to isolate an aspect or feature of the system, and evaluate its impact on data pertaining to events and timing, and successful task completion. The complexity of evaluation of interactive retrieval systems, however, lies with the complex nature of the task itself. It is likely that a user will engage in many system interactions in the course of one task (Draper & Dunlop, 1997), thus making it difficult to ascertain the point at which the interaction can be deemed to have ended and assessment of system contribution to take place. This would seem to make a strong argument for directly obtaining users' assessments of these systems based on our understanding of the search processes and tasks in which the user is engaged. Our investigation into mental models represents a step in this direction in highlighting the users' view of process as underlying the key evaluations. It would seem that the challenge of developing evaluation criteria for interactive retrieval systems is at the point where research in information seeking behaviour models and research in retrieval system development begin to come together.

References

Brandt, D. S., & Uden, L. (2003). "Insight into mental models of novice Internet searchers", *Communications of the ACM*, 46(7), pp. 133-136.

Burke, M. (2001). "The use of repertory grids to develop a user-driven classification of a collection of digitized photographs", *Proceedings of the 64th ASIST annual meeting, Washington* (pp. 76-92). Medford, NJ: Information Today, Inc.

Corbridge, C., Rugg, G., Major, N. P., Shadbolt, N. R., & Burton, A. M. (1994). "Laddering: technique and tool use in knowledge acquisition", *Knowledge Acquisition*, 6, pp. 315-341.

Crudge, S. E., & Johnson, F. C. (2004). "Using the information seeker to elicit construct models for search engine evaluation", *Journal of the American Society for Information Science and Technology,* 55(9), pp. 794-806.

DeKleer, J., & Brown, J. S. (1983). "Assumptions and ambiguities in mechanistics mental models". In: Gentner, D., & Stevens, A. L. eds. *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 15-34.

Dillon, A. (2004). *Designing usable electronic text: ergonomic aspects of human information usage.* 2nd Edition. London: CRC Press.

Dillon, A., & McKnight, C. (1990). "Towards a classification of text types: a repertory grid approach", *International Journal of Man-Machine Studies*, 33(6), pp. 623-636.

Draper, S.W., & Dunlop, M.D. (1997). "New IR – New evaluation: the impact of interactive multimedia on information retrieval and its evaluation", *The New Review of Hypermedia and Multimedia*, 3, pp.107-122.

Dunn, W. N. (1986). "The policy grid: a cognitive methodology for assessing policy dynamics". In: Dunn, W. N. (ed.) *Policy analysis: perspectives, concepts and methods.* Greenwich, USA: JAI Press, pp.355-375.

Fetzer, J. H. (1993). "The argument for mental models is unsound", *Behavioral and Brain Sciences,* 16(2), 347-348.

Fransella, F., Bell, R. and Bannister, D. (2003) *A Manual For Repertory Grid Technique*. John Wiley & Sons

Hassenzahl, M., & Trautmann, T. (2001). "Analysis of web sites with the repertory grid technique". Retrieved August 2003 from http://www.tu-darmstadt.de/fb/fb3/psy/soz/veroeffentlichungen_mh/ Chi01_hass_rgt.pdf

Hinkle, D. (1965). *The change of personal constructs from the view point of a theory of construct implications.* Unpublished Ph.D. thesis, Ohio State University.

Holscher, C., & Strube, G. (2000). "Web search behaviour of Internet experts and newbies"; *Computer networks*, 33, pp. 337-346.

Hunter, M. G. (1997). "The use of RepGrids to gather interview data about information systems analysts", *Information systems journal*, 7, pp. 67-81.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). "Real life, real users, and real queries: a study and analysis of user queries on the Web", *Information processing and management,* 36(2), pp. 207-227.

Johnson, F. C., Griffiths, J. R., & Hartley, R. J. (2001). *DEVISE: a framework for the evaluation of Internet search engines.* Library and Information Commission Research Report 100.

Johnson, F. C., Griffiths, J. R., & Hartley, R. J. (2003). "Task dimensions of user evaluations of information retrieval systems", *Information research*, 8(4), Retrieved July 9, 2005 from http://informationr.net/ir/8-4/paper157.html

Johnson-Laird, P. N. & Byrne, R. M. (1991). *Deduction.* Hillsdale, NJ: Lawrence Erlbaum.

Kelly, G.A. (1991). *The psychology of personal constructs*. London: Routledge (Original work published 1955).

Landfield, A. W. (1971). *Personal construct systems in psychotherapy*. Chicago: Rand McNally.

McKnight, C. (2000). "The personal construction of information space", *Journal of the American society for information science*, 51(8), pp. 730-733.

Moukdad, H., & Large, A. (2001). "Users' perceptions of the Web as revealed by transaction log analysis", *Online information review*, 25(6), pp. 349-358.

Moynihan, T. (1996). “An inventory of personal constructs for information systems project risk researchers”, *Journal of information technology*, 11, pp. 359-371.

Muhr, T. (1997*) Atlas/ti: short user’s manual*. Berlin: Scientific Software Development.

Muramatsu, J., & Pratt, W. (2001). “Transparent queries: investigation users’ mental models of search engines”, *ACM SIGIR*, New Orleans, Louisiana, USA, September 9-12, 2001.

Norman, D. A. (1983). “Some observations on mental models”. In: Gentner, D. & Stevens, A. L. eds. *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 15-34.

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

Ratzan, L. (2000). “Making sense of the Web: a metaphorical approach”, *Information research,* 6(1). Retrieved July 9, 2005 from http://informationr.net/ir/6-1/paper85.html

Reynolds, T. J., & Gutman, J. (1988). “Laddering theory, method, analysis and interpretation”, *Journal of advertising research*, 28, pp.11-31.

Rugg, B., Eva, M., Mahmood, A., Rehman, N., Andrews, S., & Davies, S. (1999). Eliciting information about organisational culture via laddering. *Proceedings of the Enterprise Management and Resource Planning Studis*, San Salvador, Venice, November 25-26. Retrieved August 2003 from http://leks.iasi.rm.cnr.it/emrps’99/papers/rugg_et_al.pdf

Saracevic, T. (1996). Modelling interaction in information retrieval: a review and proposal. *Proceedings of the 59th annual ASIS meeting*, Baltimore, October 21-24, (pp. 3-9). Medford, NJ: Information Today, Inc.

Saracevic, T. (1997). “Extension and application of the stratified model of information retrieval interaction”. *Proceedings of the annual meeting of the American Society for Information Science,* 34, pp. 3-9

Seadle, M. (2003). Editorial: mental models for personal digital assistants (PDAs). *Library high tech*, 21(4), pp. 390-392.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). “Analysis of a very large Web search engine query log”, *SIGIR Forum*, 33(1), 6–12.

Slone, D. J. (2002). “The influence of mental models and goals on search patterns during Web interaction”,*Journal of the American society for information science and technology,* 53(13), pp.1152-1169.

Spink, A. (2002). “A user-centred approach to evaluating human interaction with web search engines: an exploratory study”, *Information processing and management,* 38(3), pp. 401-426.

Spink, A., Bateman, J., & Jansen, B. J. (1999). “Searching the Web: a survey of Excite users”, *Internet research*, 9(2), pp. 117-128.

Spink, A., Jansen, B. J., & Ozmultu, H. C. (2000). “Use of query reformulation and relevance feedback by Excite users”,*Internet research*, 10(4), pp. 317-328.

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). “Searching the web: the public and their queries”, *Journal of the American society for information science and technology*, 52(3), pp. 226-234.

Staggers, N. & Norcio, A. F. (1993). “Mental models: concepts for human-computer interaction research”, *International journal of man-machine studies*, 38, pp. 587-605.

Stewart, V. & Stewart, A. (1981). *Business applications of repertory grid*. London: McGraw-Hill.

Strauss, A. & Corbin, J. (1998). *Basics of qualitative research: techniques and procedures for developing grounded theory* (2nd ed.). London: Sage.

Su, L. T. (2003). “A comprehensive and systematic model of user evaluation of Web search engines: I. Theory and background”,*Journal of the American society for information science and technology*, 54(13), pp. 1175-1192.

Sullivan, D. (2003). Nielsen NetRatings Search Engine Ratings. Retrieved July 9, 2004 from http://www.searchenginewatch.com/reports/article.php/2156451

Tan, F. B. & Hunter, M. G. (2002). “The repertory grid technique: a method for the study of cognition in information systems”, *MIS Quarterly*, 26(1), pp. 39-57.

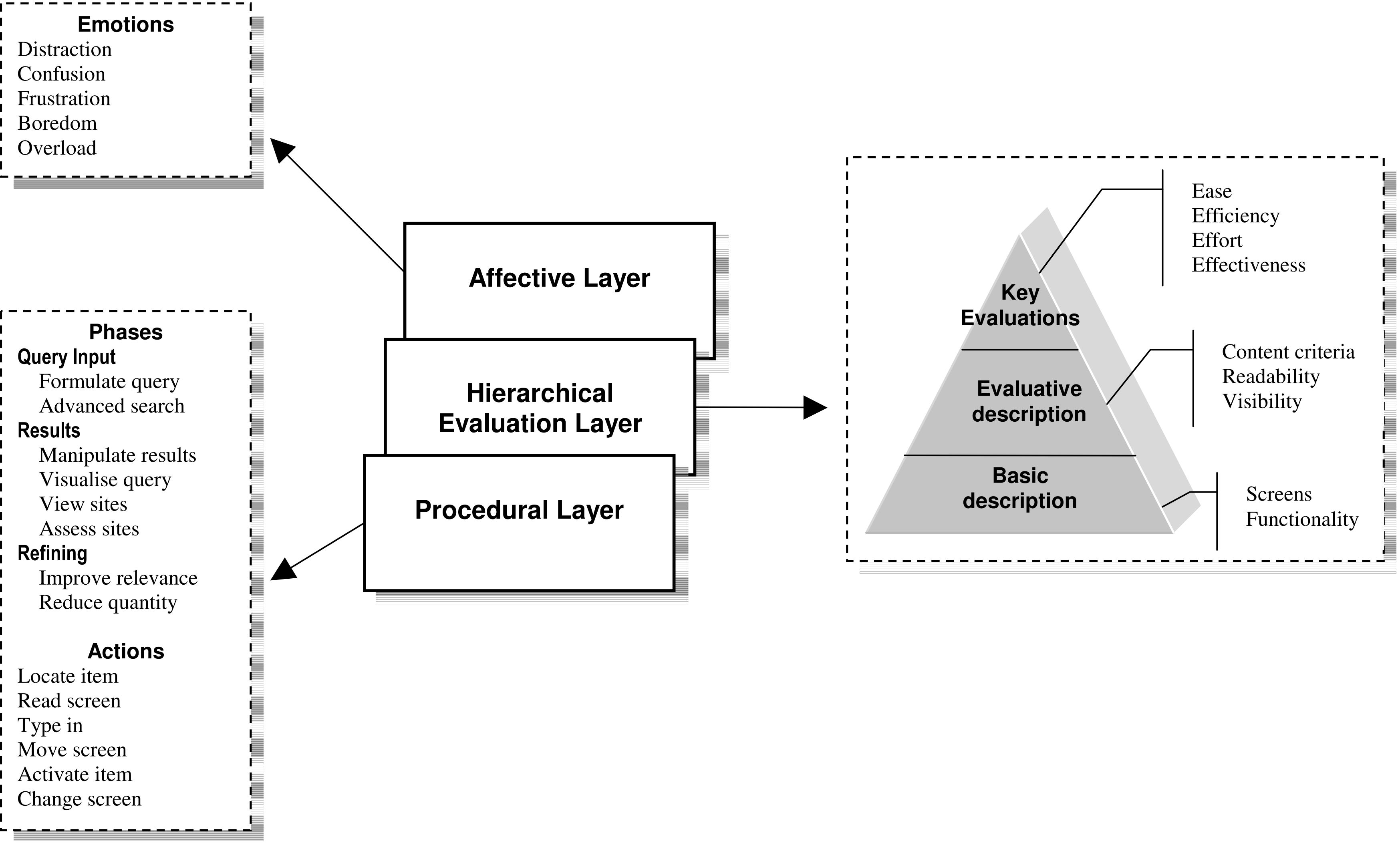
Thatcher, A. & Greyling, M. (1998). “Mental models of the Internet”, *International journal of industrial ergonomics,* 22, pp. 299-305.

Van der Veer, G. C. (1989). “Individual differences and the user interface”*’Ergonomics,* 32(11), 1431-1449.

Whyte, G. & Bytheway, A. (1996). “Factors affecting information systems' success”, *International journal of service industry management*, 7(1), pp. 74-93.

Xie, H. (2003). “Supporting ease-of-use and user control: Desired features and structure of Web-based online IR systems”, *Information Processing & Management*, 39(6), pp. 899-922.

Zhang, X. & Chignell, M. (2001). “Assessment of the effects of user characteristics on mental models of information retrieval systems”, *Journal of the American society for information science and technology*, 52(6), pp. 445-459.


Emotions
Distraction
Confusion
Frustration
Boredom
Overload
Phases
Query Input
Formulate query
Advanced search
Results
Manipulate results
Visualise query
View sites
Assess sites
Refining
Improve relevance
Reduce quantity
Actions
Locate item
Read screen
Type in
Move screen
Activate item
Change screen
Affective Layer
Hierarchical Evaluation Layer
Procedural Layer
Key Evaluations
Evaluative description
Basic description
Ease
Efficiency
Effort
Effectiveness
Content criteria
Readability
Visibility
Screens
Functionality