Chapter 1

1.1 Introduction and rationale

Search engines are now recognised as the dominant information seeking tool utilised by people using the WWW (Jansen & Spink, 2006; Slone, 2002). A number of studies interested in the information retrieving methods employed by users have long recognised this (Slone, 2002). As such, there has been a gradual shift from the focus of usersô mental models of library databases (Borgman, 1986) to a more active interest in the kind of mental models users develop regarding search engines (Jansen & Spink, 2006; Muramatsu & Pratt, 2001).

This interest in mental models and search engines has perhaps been primarily motivated by two factors; to establish whether search engines are designed in a manner that allows for an effective search user interface (Spink, Jansen, Blakely & Koshman, 2006; Vaughn & Resnick, 2006); and to establish what kind of role mental models play in the way in which search engines are utilised by users and the implications thereof (Slone, 2002; Muramatsu & Pratt, 2001).

Mental models have long been seen as difficult to define (Staggers & Norcio, 1993) however, a generic definition outlines a mental model as the mental representation of an external event or as being an analogous mental picture or abstraction of particular external happenings (Borgman, 1986; Thatcher, 2003). Mental models are developed through the usersô interactions with a given system or through the observation of another person using the system (Borgman, 1986). In the context of search engines the mental model would refer to individualsô mental ñpicturesò of the search engine system with which they interact. Staggers and Norcio (1993) highlight the ability of mental models to predict future events as well as the explanatory and descriptive power they provide. It should be clear, problematic as this concept may be said to be, why an understanding of mental models in the realm of the most dominant method of retrieving information on the WWW (search engines) has been so important (Spink et al., 2006).

Conceptual models ñ[are] invented by teachers, designers or scientists as accurate, consistent and complete representation[s] of the target systemò (Staggers & Norcio, 1993, p. 588). Theoretically then, where relevant, there should be good alignment between usersô mental models and the designersô conceptual models if effective use of a system is to be observed. Staggers and Norcio

(1993) also note that greater exposure to a system should improve the quality of the mental model and hence its alignment with the conceptual model.

Studies that have looked at mental models of search engines have either done so very briefly or inferred mental models indirectly from log-file data on user formulated queries (Thatcher & Greyling, 2003). Spink et al. (2006) found that usersô query formulations indicated a poor understanding of the manner in which search engines work. Similar results have been found by researchers like Muramatsu and Pratt (2001), Slone (2002) and Spink, Jansen and Ozmultu (2000) with similar query-based analyses. Zhang (2008) found undergraduatesô mental models of search engines also varied in complexity and accuracy with most users showing fairly naive mental models. In short, the general sense is that usersô interactions with search engines indicate a poor understanding of how they work and hence a poor mental model is inferred (Liaw & Huang, 2006).

In an attempt to provide a more direct approach to the investigation of mental models of search engines, Thatcher and Greyling (2003) developed and put participants through a set of defined tasks that they argued gave a good indication of their mental models of search engines. The results again indicated that mental models of search engines were in fact quite incomplete and inaccurate based on performance and information retrieved from participants to formulate/derive their mental models (Thatcher & Greyling, 2003). The study is unique in that it considered multiple search engines, previous studies have generally focused on one search engine at a time (Spink et al., 2006), and actively asked users for their description of the system used (Thatcher, 2003).

The results are similar to those of previous studies however. This would suggest that users develop inaccurate and incomplete mental models during their interactions with search engines. One reason for this might be that the systems (search engines) that users are exposed to and use to develop their mental models are poorly designed (Rose, 2006; Thatcher & Greyling, 2003). It might also be because usersô retrieve information adequately enough without fully understanding or wanting to understand the system with which they interact (Note that most of the studies done, being based on log-file data, do not actually know if users succeeded in their queries or not) well enough.

Staggers and Norcio (1993) noted how mental models of any given system should improve with greater system exposure. As such, this research study seeks to explore whether mental models and search engine systems have aligned better over time. Applying the same methodology used by Thatcher and Greyling (2003), with data Thatcher (2003) collected in 2000, the study aims to determine the kind of role time might play in the alignment of mental models and search engine systems. As no longitudinal research has been conducted on the quality of mental models of search engines and because the exposure people have to the internet could be said to have increased quite significantly since the Thatcher and Greyling (2003) investigation (Speckman, 2010), a meaningful platform is set to assess whether the changes since then have seen a closer alignment of mental models with conceptual models and search engines.

1.2 LITERATURE REVIEW

1.2.1 What is a mental model?

It is a seemingly difficult task to both find and provide a formulaic definition for mental models that satisfies authors with an interest in this area (Doyle & Ford, 1999; Staggers and Norcio, 1993; Rutherford & Wilson, 1992). This could be because the concept of a mental model has been applied in a number of different fields and as such, different authors emphasise different aspects (Staggers and Norcio, 1993). The conceptualisation of what a mental model is has spanned the fields of cognitive psychology and ergonomics, engineering, engineering psychology, philosophy and many other fields (Staggers & Norcio, 1993; Thatcher & Greyling, 2003).

Apart from the problem of defining mental models, there was also the issue of multiple-term-use when considering mental models (Spicer, 1998; Staggers & Norcio, 1993). A number of authors, including Norman (1988), use the term óconceptual modelsô interchangeably with that of ómental modelsô. Hollnagel (1988), when writing about mental models, used the term óoperator modelô. Clarity in this regard came through the distinction made by, among others, Staggers and Norcio (1993), who defined conceptual models as the description and understanding provided by a designer, expert or teacher of the system in question. To make the distinction even clearer, Staggers and Norcio (1993) noted how designers, in all likelihood, needed to have clear mental models of usersô mental models of their systems in order to effectively design and improve their systems. Also, accurate conceptual models of those systems would require a good understanding of the system in question from multiple perspectives (users, fellow experts etc). This definitional distinction, unlike that of mental models, is quite clear and leaves practically no doubts as it clearly articulates that conceptual models are descriptions provided by designers, experts, and teachers about a given system (Spicer, 1998; Staggers & Norcio, 1993; Thatcher & Greyling, 2003). Providing the different descriptions and definitions of what a mental model is will allow for a generic definition to be provided and it is the understanding gained through this consequent definition that will carry the research forward with a good level of clarity.

According to Borgman (1986), a mental model is ña model of the system that the user builds in his or her mindò (p. 268). Johnson-Laird (1983) saw a mental model as being a small-scale mental representation of the kind of reality users are faced with. Indeed, the author credited with establishing the concept of a mental model, Craik (1943), proposed that the key to understanding the world around us rests on our ability to create symbolic representations of it internally through óworking modelsô of sorts. These óworking modelsô of the world around us are what Craik (1943) contended to be mental models.

These initial conceptualisations of a mental model are informative and provided much of the groundwork for work in this area but they lacked quite a bit of detail as to what exactly mental models consist of (their true content-nature) and how exactly mental models are formed (Staggers and Norcio, 1993). In fact, Thatcher (2003) noted how Johnson-Lairdôs (1989) revised definition, which incorporated the role spatial perception played in the formulation of mental models, was considered as being far too broad to have any relevant real-world applications in terms of a number of things including system conceptualisations and design. More definitions seeking to provide clarity have since been provided.

Booth and Brown (1989) saw mental models as metaphors that provide analogy-based methods of learning. That is, ñpictorialò mental representations that provide analogous content that allows for a basic understanding of specified parts of the external world. Mental models have also been described as abstract, cognitive mappings that assist in the understanding of spatial and relational properties in the physical world and, importantly, cyberspace as well (Hirtle & Jonides, 1985;

Liben, 2001; Thatcher & Greyling, 2003). Borgman (1986) noted that mental models are in fact developed through the userọs interaction with a system or through their observation of somebody else using that system. So, direct or indirect interaction with the target system is a key condition for the formulation of a mental model (Borgman, 1986; Thatcher & Greyling, 2003). Thatcher and Greyling (2003) noted that the fact that mental models can also be acquired indirectly has a number of implications. Firstly, that knowledge (mental model) transferability from one system to another is a real possibility; and that users may experience conflict between new mental models and those transferred from another system when interacting with a new system that is different to the older/other one. So critically, interaction, which is a condition for mental model formulation, can contribute towards distorted and unsuitable mental models (Spicer, 1998; Staggers & Norcio, 1993; Thatcher & Greyling, 2003). This point touches on another key quality of mental models; namely their tendency to be incomplete in form.

A thorough consideration, through an examination of both the theoretical content and empirical studies testing mental models in some or other way by authors and researchers like Borgman (1986), Slone (2002), and Staggers and Norcio (1993) has resulted in another point being made about mental models. Specifically, the argument which states mental models seem to be incomplete and can never really become totally complete. Also, Thatcher and Greyling (2003) note how the level of accuracy mental models are believed to have, particularly in complex problem domains of computer systems, is seen as being quite poor (Staggers & Norcio, 1993). Indeed, the very definition of what a mental model is remains something of a well understood mystery (Borgman, 1986; Senge, 1990; Spicer, 1998; Staggers & Norcio, 1993).Notwithstanding all of these issues, a satisfactory definition can be provided. It is quite clear, at a generic level at least, that a mental model is a mental representation of an external event; or can be seen as being an analogous mental ñpictureò or abstraction of particular external happenings that forms and (continues to) develops upon interaction with or secondary (observational) exposure to the external system/event in question.

<u>1.2.2 Why are mental models important?</u>

The importance of mental models is quite evident even if the manner in which mental models are defined seems broad and convoluted. If Craikôs (1943) account of mental models is even remotely accurate, then the relevance of mental models could almost be seen as essential for effective human functioning. Craik (1943) held that in order for people to make sense of the world around them they need to represent the world around them symbolically. Mental models provide the means Craik (1943) and many others (e.g. Johnson-Laird, 1983; 1989) proposed people use to understand the world around them. *Understanding*, logically speaking, is a fundamental aspect of potentially effective functioning in any context. Mental models then, could be said to be fundamental and necessary for every human being as they allow for significant, and perhaps more importantly, meaningful interactions with the external world. It must be clarified that while the reasoning above is justified, it must be read as the strongest possible interpretation of the instrumental potency of mental models from a theoretical perspective. While this interpretation is not misleading, it is certainly not completely true either. This will be illustrated with examples of research done in this area in later sections (Doyle & Ford, 1999; Spicer, 1998).

In the context of a complex and dynamic technological environment that is characterised by complex computer systems, system comprehension, both on an individual and organisational level, becomes especially important (Morgan, 1997; Senge, 1990). The importance of mental models in this regard could be seen as quite significant. According to Borgman (1986), mental models have the potential to facilitate effective usage of technological equipment such as computers and enhance information retrieval processes in complex database environments. Kerr (1990) argued for the importance of mental models over prompted navigational cues in effective navigational functioning through intricate systems. The efficacy of ówayfindingô through a very complicated database could be seen as being reliant on the kind of cognitive maps users have at their disposal and in that sense, the cognitive maps users have could be seen as only being as good as the environment in which users find themselves and develop these maps (Hodkinson, Kiel, & McColl-Kennedy, 2000; Kerr, 1990). This point leads to another salient aspect of mental models in the context of complex computer and database domains; that of their role in system design.

Rouse and Morris (1986) saw mental models as having the ability to describe, explain and predict events in the relevant system domain. In this regard, the value of mental models in complex computer domains could be seen as taking on even greater significance. Not only would their importance be crucial in the ability to enhance usersô effective use of systems on a number of levels, but also, as pointed out by Rouse and Morris (1986), from the perspective of system design, mental models could possibly assist in the estimation of a systemôs state. That is, the manner in which peoplesô mental models help them work their way in and through a system, could provide valuable information as to the quality of the design of that given system.

In a sense then, the cognitive representations (mental models) of a system that Scaife and Rogers (1996) see as an essential aspect of effective interaction between users and systems, could be viewed as providing a means to inform system developers and help them to develop not only effective systems but effective conceptual models of those systems as well (Rouse & Morris, 1986). Rouse and Morris (1984) also highlighted a reason more acutely related to design issues as concerning mental models. They argued that given the complex state of systems and most likely, the even more complex state future systems are likely to have, different mental models for different aspects of the same system are also likely to develop. Thatcher and Greyling (2003) note that if this statement holds, the implications are that it is quite possible for users to be adept at using certain aspects of a system while remaining relatively poor users of a system in its entirety. Indeed, Thatcher (2003) made the point that experienced users and novices could quite possibly perform in a similar manner on certain information retrieval tasks as the nature of those tasks may lend themselves to more intuitive and less complicated mental model formulations and use.

In short, the importance of mental models is significant if the accounts given above are taken into consideration. Not only are mental models considered to be important for ósense-makingô and understanding in complex environments (i.e. cyberspace and the physical world) but they could also be seen as an important means which system evaluations and design could use to improve on their current states (Kerr, 1990; Rouse & Morris, 1986; Rouse & Morris, 1984; Thatcher, 2003; Thatcher & Greyling, 2003; Zhang, 2010).

1.2.3 How do mental models work?

It seems sensible to infer *how* mental models work from the definition of *what* they are (Staggers & Norcio, 1993). Only, the definition of mental models has been quite problematic in many respects (Staggers & Norcio, 1993; Spicer, 1998) as scholars in different disciplines have tended to place emphasis on different aspects. This does not mean a reasonable account of how mental models work is not possible though. Thatcher and Greyling (2003) point out that mental models form upon interaction with a system or observation of someone else using a system. Borgman (1986) noted this too as the conclusion she reached about system engagement, directly and/or indirectly, leading to mental model formulation. That is the first thing to note. After a mental model is formed within a particular context it gradually develops as the user is exposed to that system or environment. Senge (1990) notes how mental models use a variety of ñif x then yò combinations in their attempt to describe and explain a system. Upon a satisfactory state of comprehension usersô mental models allow for effective predictions of the manner in which a system will respond (Rouse & Morris, 1986).

When an individual interacts with a particular system they form ideas about how that system works. The more they learn about the system the more likely it is that they will figure out how that system works and responds to certain commands, actions or behaviours. In figuring out how the system works individuals will be able to anticipate system responses prior to even engaging the system or giving it the relevant commands. This will mean that the user is able to have ideas that are capable of describing, explaining and predicting system behaviour and its functional properties (Johnson-Laird, 1989; Rouse & Morris, 1986). That is, the user will be able to *understand* how a given system works. Experience or knowledge of other/older systems can also help users in understanding the new systems they are faced with. This past experience may though, also hinder the process of understanding new systems (Staggers & Norcio, 2003).

To bring the abstract concept of mental models ñdown to earthò the example of a jungle-gym can be used. Upon initial interaction with a jungle-gym the user, assumed to be a child here, learns about the net used to climb up the jungle-gym as well as the wooden-platform at the top, the slide, and the beam pole that can also be used to go back to the ground. Aspects of how much force and speed can be used in climbing the net, going down the slide and beam pole respectively, and various other things are developed the more interactive the user becomes.

Trivial as it may seem for an adult (likely expert), the child will then be able to describe how to use all the different parts of the contraption, explain when optimal enjoyment is likely to be achieved and even predict when injuries are likely to occur. That is to say, the child will have developed a thorough understanding of how the system works. And it must be noted that this is possible from the child simply observing other children using the system. Also, their previous knowledge of certain aspects of the system (e.g. the safety net) could also assist or perhaps hamper their mastery of the new system. This analogy provides an apt, substantially less abstract description of how mental models work and further illustrates just how important mental models could be seen as being.

1.2.4 The change of mental models over time

It is a widely held notion that mental models develop as a result of exposure to a system, directly and/or indirectly (Borgman, 1986; Thatcher & Greyling, 2003). As a user continues to be exposed to a system, more things about that particular systemҧs operational functions are likely to be revealed. The formulation and development of mental models through interactive means relating to a particular operational system is a basic premise of the theoretical rhetoric on mental models (Rouse & Morris, 1986; Thatcher & Greyling, 2003). While the manner in which mental models are formed is simple enough to comprehend, a number of issues seem to arise from the idea of mental model development, particularly in a complex computer system environment (Spink et al., 2000; Staggers & Norcio, 1993). Considering that a number of authors believe that mental models in environments of this complex nature are generally inaccurate, incomplete and poorly formed, the development of mental models in this context could be seen as quite puzzling (Slone, 2002; Spink et al., 2000; Thatcher & Greyling, 2003). The main reason why this may be puzzling is because it is both implicitly clear and explicitly stated that mental models probably improve in accuracy and completeness over time and result in better performance with greater use of and exposure to a system (Staggers & Norcio, 1993; Doyle & Ford 1999).

The idea here is that the more a user is exposed to a particular system the better the opportunity for mental models to become more accurate and complete. So the greater the time a person spends using a particular system the better they become at using it, which could imply healthy mental model development of that system. Indeed, Slone (2002) found that those users who were most motivated to search for information on the WWW were the most effective information finders when compared to those whose motivation was not as high. The point here is not about motivation but about system exposure. It is not surprising that those people Slone (2002) found to be most proficient in using search engine systems also happened to be the people who spent the greatest time interacting with those systems. That is, Sloneǿs (2002) results provide evidence of the importance of system exposure and, by implication, the role time plays in the development of mental models as relating to the relevant system. Muramatsu and Pratt (2001) found similar trends when investigating the kind of mental models users have of search engines. Admittedly, Muramatsu and Pratt (2001) found that usersô mental models of search engines were generally quite poor, but they noted and inferred from their data that a noteworthy difference in the quality of query formulations of those users who spent more time using the search systems existed when compared to those who were deemed less dedicated or inclined.

Theoretical inferences from the generic definition of mental models lead to the conclusion that mental models develop over time and these findings in the two investigations provide empirical evidence for this idea. As stated earlier, it is a clear implication, implicit and explicit, of the theoretical content on mental models that they are likely to improve over time (Borgman, 1986). It is worth noting that improvement here is not and cannot be necessarily aligned to system proficiency, but only to differences found in usersô differing uses of a system. This is a particularly salient point in the domain of complex computer systems as most of the evidence on the quality of mental models in this environment points towards mental models of a quite poor nature (Spink et al., 2000; Slone, 2002; Thatcher, 2003; Thatcher & Greyling, 2003).

Studies conducted on the nature of mental models by researchers like Slone (2002), Muramastu and Pratt (2001), and Zhang (2008; 2010) have all been cross-sectional in design. These studies have each pointed towards the largely inaccurate and incomplete mental models of search engines people have. To compensate for system exposure these studies have looked at participantsô experience with a particular system. Given that the theoretical assertions about mental models suggest that greater system exposure over a sustained period of time contributes to the effective development of mental models, more longitudinal studies provide more meaningful knowledge in this regard. While all these studies have tried to account for system exposure by measuring the amount of exposure users have had, the most effective way to test the theory concerning how mental models develop might require a more longitudinal approach.

Indeed, it is surprising that there are almost no studies that have attempted to apply a longitudinal design to track the development of mental models over a sustained period of time. One of the very few studies to employ a longitudinal design to track mental models was conducted by Oleson et al (2010). In their study Oleson et al (2010) assessed the mental models of two groups of grade 1, 3 and 5 had of computers. The first group of children was interviewed in 1999 and the second group was interviewed in 2009. The findings indicated that while the mental models of both groups were equally well formed, children in 2009 were developing their mental models of computers at a much younger age. Oleson et al (2010) also noted how the analogies of computers differed among the two groups with the 2009 group showing more progressive analogies. This longitudinal study gives an indication of the kind of information that can be extracted about mental models if a longitudinal design is employed. It also highlights the fact that studies that assess mental models in the context of a cross-sectional design are probably failing to capture important information about mental models.

The idea of mental models improving over time is also interpretable in another sense. As mental models form in contexts permitting interaction it would be irresponsible to neglect the context in which the interaction occurs to understand its impact (Craik, 1943). That is, given that environments change, and perhaps none more so than the technological environment, this has a number of implications for system exposure. Simply, because technology develops so rapidly, what was exclusive a few months ago may turn out to be readily available to many now (Speckman, 2010). The growth in ownership of the personal computer and the increased access to the internet; as well as the speed of the internet because of increased bandwidth, are examples of this phenomenon. The adequate analysis of peoplesô mental models becomes, on some levels, more difficult when comparisons between users separated by a decade are considered, given the fast changing environment. The analysis can also be a more meaningful endeavour when there are more people using systems more of the time as this provides greater variability in system usage and greater scope for interpreting results meaningfully (as evidenced in the results of Oleson et al., 2010). This is so even if comparatively, there might be contextually based differences (e.g. technological capability of U.S.A compared to R.S.A). This is the sense in which the context can, and does in this instance, play a role in both the type of interactions that occur and the magnitude of the interactions. This is a point worth noting in the light of the questions that will be asked and the general nature of the current study.

1.2.5 Search engines

Search engines have become the primary method of looking for information on the WWW (Jansen & Spink, 2006; Slone, 2002; Spink et al., 2006). While user-managed knowledge databases like Wikipedia have also risen in popularity, they too are often reached through search engines (Sullivan, 2006; Wilkinson & Huberman, 2007). In fact, a number of people rely solely on the internet to find information and they view search engines as the only way to access the internet (Graham & Metaxas, 2003; Koshman, Spink & Jansen, 2006; Muramatsu & Pratt, 2001). As Thatcher and Greyling (2003) pointed out, the fact that people talk about ñGoogling itò when referring to a means of searching for information indicates the central importance search engines have taken on since their inception in the early 1990ôs (Spink et al., 2006). Search engines could almost be seen as the ógate-keepersô to the WWW as a large proportion of the population worldwide knows no other means of accessing information other than through a particular search engine or a number of different search engines (Spink et al., 2006; Wang, Hawk, & Tenopir, 2000; White & Livonen 2001). The importance of search engines to people in the information retrieval process is undoubtedly fundamental. The era of research on user aspects concerning library databases is almost over or has become, at the very least, secondary (Jansen & Spink, 2006; Muramatsu & Pratt, 2001; Thatcher & Greyling, 2003).

The mechanics of search engines vary quite significantly and the consequent result is that different search engines produce different outcomes concerning the same user queries; and their search performances are qualitatively different because of this (Bharat & Broder, 1998; Gordon & Pathak, 1999; Spink et al., 2006). All search engines are in fact databases (Spink et al., 2006). In order to build their databases search engines use ócrawlersô. Crawlers are computer programmes built to scan and retrieve information from the Web (or information related to it) that they are programmed to find (Gordon & Pathak, 1999). This information is then stored in the search engineôs database. Search engines then apply a set of mathematical algorithms that have been programmed into the system in order to search their own databases for information that matches the search terms requested from a user. A separate set of algorithms is then applied by search engines to systematically rank the results yielded for a particular search query in order of importance or relevance (related to the userӳ search terms and the search engines specific ranking algorithm). This particular function is dependent on the type of algorithms used by a given search engine in defining ԲelevanceԠand means varyingly ordered results may be produced for identical queries when using different search engines. The variability in results would also be influenced by the manner in which crawlers are programmed as different Web crawlers could result in different search engines searching the Web in different ways and producing varying databases.

As different search engines use different algorithms and subsequent indexing techniques to sift through and rank the information they have built up in their databases in order to respond to queries, the question of which search engine is better seems relevant (Spink et al., 2006; Thatcher & Greyling, 2003). However, as Spink et al. (2006), Thatcher and Greyling (2003) and Mowshowitz and Kawaguchi (2005) have noted, it really is a question of each search engine systemӳ operational set-up being better suited to certain queries rather than others. Indeed, Spink et al. (2006) tentatively assert that the most complete option to exercise in retrieving information from the WWW may well be a metasearch engine. This is a search engine system that assesses all of the results of search engines like Yahoo, Google and MSN Search, amongst many others, and takes what it determines to be the most relevant results from each to compile its own results page for any given query (Spink et al., 2006). This seems an enticing and sensible way to search, but like search engines, different metasearch engines have unique algorithmic properties, implying that different metasearch results will emerge depending on the metasearch engine being used (Spink et al., 2006). Also, the popularity of search engines and a clear lack of knowledge or interest regarding metasearch engines mean this avenue is unknown and considerably under-used and as such, fair judgment of the quality of metasearch enginesԠprocesses is currently untenable (Spink et al., 2006)

In their study, Thatcher and Greyling (2003) found that subjects had little or no idea at all that different search engines provide different results. Even though there are elements that are similar in the way that search engines operate, the differences evident between them are also significant enough that it would be expected usersô would notice them (Spink et al., 2006). Given this fact, Thatcher and Greylingös (2003) findings were somewhat surprising. They concluded that usersô mental models of search engines were quite poor and suggested that the manner in which search engines are designed needs to perhaps lend itself to a clearer method of revealing its operational methods. Thatcher and Greyling (2003) suggested that search engine designs are such that it is unclear to users exactly how they work. While their work was published in 2003, the data collection occurred some three years before that. Indeed, some of the search engines mentioned in that particular study are no longer even functional now, while others have become dominant players in the market when very little was known about them in 2000 (e.g. Google). So there is something interesting to consider: The development and design of search engines over time since that study and the impact this has had on mental models of search engines.

1.2.6 The functional use of search engines over time

In more recent studies spearheaded by Spink et al. (2006), there is a clear indication that while search engines have become more efficient at face value, their respective modes of operation remain seemingly similar. This is simply a function of the kind of conclusions a number of researchers including Jansen and Spink (2006) reached based on the log file data of user queries formulated that were subjected to their scrutiny. While Kehoe, Pitkow, Sutton, Aggarwal and Rogerös (1998) and Kehoe, Pitkow and Rogerös (1998) user web surveys led them to conclude that web navigation remained still the biggest problem when it comes to using the web, it is unclear that this is still the case because of the more efficient, even if still quite disingenuous, nature of search engines (Thatcher & Greyling, 2003). That is, Spink et al. (2006), Koshman et al. (2006) and Montgomery and Faloutsos (2001), while inferring from their comprehensive analyses that user understandings of search engines were quite poor and that their mental models of search engines were quite poor, did not necessarily know whether users were actually finding what they were looking for. This is because the nature of their studies was primarily based on log file data that could not provide them with answers regarding whether users retrieved the information that they were looking for. There was also significant difficulty regarding participant identification because researchers had to use I.P addresses as a means of identifying a ñuniqueò user. The problem with this is that in a number of different contexts I.P addresses are limited and as such they may be randomly assigned to the same computer at different time intervals. This could mean that the same user could conceivably have different I.P. addresses (and appear to be two unique users). Conversely, multiple people could use the same computer with the same I.P. address (as in a public library). This would imply that different users could use the same I.P address and such an occurrence would seriously undermine any results reported based on this data (Spink et al., 2006). That is, if usersô performance was being measured more than once (or even at different times in the same day or week), there would be no way for the researchers to know for certain whether the I.P addresses were being used by the same unique users. More importantly, even if they were measuring performance only once, which was the case for most log file studies, there would be no way of truly knowing whether usersô were getting their desired outcomes.

Google is now the most used search engine (SEW, 2009; Sullivan, 2006). Along with this success has come an aura around the accuracy of this particular search engine, with people swearing by it and even replacing ñsearch for itò with the phrase ñGoogle itò (Thatcher & Greyling, 2003). This indicates the popularity of this particular search engine but also the fact that people seem to be getting what they want from it; to the point where they give it glowing endorsements (Sullivan, 2006). While the greater majority of research conducted on usersô proficiency in using search engines has concluded that their understanding is quite poor, and from this some have inferred that user mental models of search engines are not good, there seems to be something very appealing regarding Google in particular. The fact that most people use Google when there are many other search engines that are available would imply that Google is at the very least, more appealing than other search engines. While it is probably difficult to infer from its popularity that people are more satisfied with Google, it remains a distinct possibility that they are.

The point above is simply meant to illustrate that even though, from a design perspective, Google and various other search engine systems have seemingly done very little to help users understand how they work, they have been able to adjust their internal mechanics and usability standards to a level that does not seem to require them to reveal or even attempt to reveal their modes of operation. Resnick and Vaughanôs (2006) recommendation for search engines to appeal to the more ñinnate strategiesò (p. 777) of users seems to have been heeded by search engine designers over the last decade. Certainly, and perhaps even overriding the evidence that users display poor query formulations and mental models regarding search engines (Spink et al., 2006; Thatcher & Greyling, 2003), search engine systems seem to have been improved to a level that undermines the userõs need for and motivation to acquire sound mental models of these systems. Admittedly, search engines have remained disingenuous in revealing their operational mechanisms (Zhang, 2008). Even so, though the search enginesô interfaces have hidden many of the operational functions, like their database collection methods and the rules guiding how results are obtained and ranked among other things, they have seemingly become more user friendly. This is one of the ideas the current research wishes to explore quite carefully. If there is evidence, as usersô preferences and perceptions of search engines (in particular Google) seem to indicate, that the development of search engines has advanced to such a level as to make comprehension of the systems underlying mechanics redundant, that could imply that the subsequent need for users to develop or even want to develop effective mental models of the respective search engines is both low and deemed unnecessary. Again, sight must not be lost of the fact that this point is more about the improvements in design made by search engine systems over time than it is about usersô mental models of these search engines. That particular discussion now follows.

1.2.7 Mental models and directed search tasks using search engines

The literature on mental models of search engine systems, while significantly limited, seems unanimous in its conclusion that, in fact, users have really poor mental models of search engines (Spicer, 1998; Thatcher & Greyling, 2003). A number of methods have been used to evaluate user performance in performing search tasks. The most popular of these is the kind that uses log analyses (Spink et al., 2006; Spink, Wolfram, Jansen & Saracevic, 2001). In fact, most of the conclusions reached on the search patterns, behaviours, and abilities of users regarding search engines, have come as a direct result of the inferences made from log data analyses (Thatcher, 2003). Spink et al. (2001) correctly point out that this method of assessing user knowledge and performance leaves a number of questions unanswered and indeed unanswerable as it is limited to an archival framework that is, by its very nature, retrospective and rigid. That is to say, while conclusions have been reached pointing out poor user comprehension of search engines and the subsequent inference that users have poor mental models of search engines, the nature of the methodology behind these conclusions has cast significant enough doubt. This is because the data used has not provided the kind of information that would make the findings more concrete. For instance, while researchers were able to establish that usersô query formulation was poor, they could not effectively assess aspects like usersô performance as well as usersô satisfaction and confidence. Researchers themselves have admitted that perhaps a more adaptable method open to more dynamic and meaningful modes of enquiry needs to be taken to solidify their own processes and conclusions (Spink et al., 2001). Their focus on one search engine at a time instead of multiple search engines has also meant a number of other factors, such as usersô system differentiation abilities and satisfaction with the search process in that context, remained a mystery (Thatcher & Greyling, 2003; Spink et al., 2006).

Kerr (1990), while admittedly not using or assessing search engines, pioneered the directed search approach in investigating the manner in which users search the internet. By setting up a particular task for the subjects in his study, Kerr (1990) was then in a position to not only monitor their progress as they performed the task, but also ask them to provide a brief description of their impressions of how the search process worked that was accompanied by an illustration. Thatcher (2003) pointed out that in fact the term ñimpressionò was ineffective as it led to participants making value judgments about their experience of the search process instead of giving a clear account of how they believed the actual search process worked. Thatcher and Greyling (2003) addressed this in their research by clearly asking their participants to provide a picture and brief description of how they thought search engines worked. This process followed after Thatcher and Greyling (2003) gave their participants general and directed search tasks to perform using a search engine (or engines) of their choice. Performance on the directed search task in particular was measured using time and number of steps taken, as well as the accuracy of the answer provided (Thatcher & Greyling, 2003).

This method of assessment, described and advocated for by Baeza-Yates and Riberio-Neto (1999), is seen as overcoming some of the problems experienced by log analysis methodology in that the researchers are in a position not only to assess whether query formulation is effectively resolved but also to assess this effectively in relation to time and steps taken to task completion.

Also, they are able to question users on various issues regarding the manner in which they carried out their searches and in so doing, better understand the kind of mental models of search engines users might have (Baeza-Yates & Riberio-Neto, 1999; Kerr, 1990; Thatcher & Greyling, 2003). Past studies, such as those conducted by Muramatsu and Pratt (2001), in assessing log file data of users with varying levels of experience, found that mental models of search engines were quite incomplete and inaccurate. They noted that users who showed considerably more experience had more improved mental models than those with less experience. Slone (2002) found that users with better mental models in fact performed web searches with more efficiency and accuracy and noted how users with more complete and accurate mental models also seemed to be those with the greater levels of experience. These studies, like those performed by Spink et al (2001) and Jansen and Spink (2006) analysed user queries to judge performance.

The results that were consistently found by these researchers about the mental models users held of search engines were that they were poor. The basic line of thinking in reaching this particular conclusion was that the queries being used to search showed too many signs of a lack of understanding regarding how search engines operate. Things such as redundancy of terms in queries as well as the tendencies of users not to select the most relevant results regarding their searches were all aspects that were identified (Spink et al., 2000). Indeed, Spink et al (2000) noted how query reformulation was also decidedly poor. All of this information that was retrieved from numerous log files led researchers to conclude that it was very likely that mental models of search engines were very poor.

The conclusions found when using the directed search task method have not deviated much from those found by the log analyses (Thatcher & Greyling, 2003). The results have indicated poor user understanding of search engine systems and the illustrations provided by subjects both in Kerrôs (1990) (where ñimpressionò was interpreted inaccurately) and Thatcher and Greylingôs (2003) study also indicated poor mental models of search engines and other navigational tools on the internet. The methodology herein employed (directed search task on a search engine system) is scarce when compared to the log analyses methodology used by so many researchers (e.g. Koshman et al., 2006; Muramatsu & Pratt, 2001; Spink et al., 2001; Spink et al., 2006). Its effectiveness is considerable and it has provided researchers with a better understanding of *why* mental models of search engines are poor instead of the simple conclusion (at least when the different methodologies are juxtaposed) that they are (Baeza-Yates & Riberio-Neto, 1999; Kerr, 1990; Thatcher & Greyling, 2003).

For instance, when Zhang (2008) employed the more dynamic methodology regarding how participants structure the web and search engines in their minds (i.e. their mental models), this was extracted through semi-structured interviews and the drawings that they provided. Zhang (2010) further refined this methodology and concluded that an emotional aspect to mental models seems to be present as well. Thatcher and Greyling (2003) were also able to elicit the mental models subjects held of search engines and the data allowed them, similarly to Zhang (2008; 2010) to go beyond data based on performance and really interrogate what was happening in usersô óheadsô. Baeza-Yates and Riberio-Neto (1999) strongly recommended this more interactive and personal approach in the investigation of information retrieval as they argued that the very definition of mental models would demand an approach of a more interactive nature if the most meaningful data was to be gathered. The results of these studies have also concluded that mental models held by users are incomplete and inaccurate with a few rare exceptions. This more interactive methodology has, however, unearthed other aspects of mental models that log file analyses would have been unable to and has thus provided richer data and greater insight.

So while using directed search tasks to determine user proficiency and the quality of their mental models has actually, thus far, corroborated the findings of log data analyses-based research into user proficiency, it has been more informative and added extra credibility to the body of knowledge in this particular domain (Baeza-Yates & Riberio-Neto, 1999). It provides a good, under-utilised template for the current studyôs endeavours in investigating the state of mental models since Thatcher and Greylingôs (2003) study a decade ago.

1.3 Aim of the study

This study aims to investigate whether mental models of search engines have improved over the last ten years.

1.4 Research questions

1. Has user-performance on the directed search task improved over time; specifically regarding the number of steps taken to task completion, time taken to task completion, and the accuracy of the answer provided?
2. Have users developed more accurate and complete mental models of search engines over time?

Chapter 2

2.1 METHODS

The following chapter describes the methodology applied in answering the research questions proposed. The variables will be identified and classified accordingly and this will be followed by a description of the research design employed. The sample and its demographic characteristics will then be provided along with the procedure followed and measures used. The statistical analyses will then be discussed along with the ethical considerations.

2.2 Research design

The research design of this study was a cross-sectional, correlational, laboratory study (Whitley, 2002). This would imply that the research methodology was open to threats of ecological validity and was often quite time-consuming (Whitley, 2002). The research also had a longitudinal dimension in that data from the current sample of 2010 (also called T2 in this report) was compared statistically with data from a demographically identical sample used in 2000 (also called T1 in this report). That is, because the samples were matched on demographic descriptors, clear longitudinal measurements which allowed for the analysis of the completeness/accuracy of participantsô mental models of search engines to be assessed were taken. Because two independent samples that were matched were used, traditional problems associated with longitudinal studies like participant attrition were avoided. As this research was correlational, no causal conclusions could be made.

Classification of variables

The following variables were statistically analysed.

- Mental model of search engine = Independent variable (regarding its impact on performance specifically)
- Time = Independent variable (as ten years have passed, time is an ex post facto IV in the study)

- Number of steps taken to task completion = Dependent variable
- Time taken on directed task = Dependent variable
- Accuracy in completing the task = Dependent variable
- Mental model completeness/accuracy of search engine = Dependent variable

2.3 Sample

The sample of this study was similar in nature to that of Thatcher and Greyling (2003). That is, it contained a good mixture of professionals (i.e. 11 academics, 4 management consultants, a medical doctor, charted accountant, lawyer, 5 I.T technicians, a senior marketing manager and a civil engineer), semi-professionals (i.e. a bookstore manager, driver, administrator and salesman), and scholars and students (i.e. 8 high school pupils, 21 undergraduates, and 22 post-graduates). This meant that a sample of 80 was attained, just as in the sample from 2000. The criterion for participation in both samples was the use of the WWW at least once.

While the original sample used a non-probability sample technique known as snowballing (Whitley, 2002), the researcherôs efforts to achieve a sample similar to that meant that the present studyôs sampling strategy was a purposive sampling technique, and it too was a non-probability sampling technique (Whitley, 2002). There was also some snowball sampling as participants were asked to recommend other potential participants who matched the requirements of the study. The sample consisted of 48 males and 32 females. Fifty of the participants spoke English as their first language, 12 spoke isiZulu, 2 spoke Sesotho, 3 spoke Afrikaans, 3 spoke isiXhosa, 3 spoke Setswana, 1 spoke Sepedi, 1spoke siSwati, 2 spoke Shona, 1 spoke Mandarin, 1 spoke Yoruba, and 1 spoke Ndebele. All the participants were fluent in English though. The sample had a mean age of 25.78 with a standard deviation of 8.7 and a range of between 14 and 60. As the sample strategy was non-probability in nature, generalisations to the greater population should be made with caution (Whitley, 2002).

Table 1: Demographic descriptors comparison

Demographic variables	2000	2010

Male	50	48
Female	30	32
Mean Age	23.28	25.78
English	51	50
South African Languages	25	24
Foreign African Languages	2	5
Euro/Asia	2	1
I.T Professionals	7	5
Engineering	1	1
Finance, Law	1	2
Academic	7	11
Management, Marketing, Medical doctor	1	7
Semi-professionals	10	5
Undergraduates	21	21
Post-graduates	25	22
Students	7	8

Table 2: WWW Experience comparison

Experience	2000 Sample	2010 Sample	t-value
Mean length of WWW use in	36.86	119.42	-14.82**

months			
Mean weekly hours on WWW	7.35	16.44	-5.49**
Mean frequency of WWW use (Never (1) once or twice a year (2); month (3); week (4); everyday (5)	4.37	4.80	-4.45**
Mean of self-rated WWW capability	3.32	3.70	-2.86*

** $p < 0.01$; * $p < 0.05$

2.4 Procedure

The researcher approached students, scholars, professionals and semi-professionals, and invited them to voluntarily participate in the study. The nature of their participation was similar to that required by Thatcher and Greyling (2003) in their study and the participant information sheet the potential participants were given outlined exactly what was expected if they chose to partake in the research (see Appendix 1).

The researcher also talked potential participants through the expectations and addressed all the questions they had. Once the people approached agreed to participate, upon their arrival at the location where the data was being collected and following their completion of the biographical (see Appendix 4) and composite WWW experience questionnaire (see Appendix 3), they were asked to complete the directed search task on a computer set up by the researcher while having their search actions recorded by onscreen capture software as they did so. Because there was a possibility that some participants (expert users or subject experts) would not use a search engine for the primary task, a secondary task (i.e. general purpose browsing task) requiring them to use a search engine was put in place. Similarly to the 2000 sample, only two participants did not use a search engine for the directed search task and required use of the secondary task. In these two cases the directed search task was used for the performance measure and the general purpose task was used to help establish participantsô mental models. The primary, directed search task was: ñFind Bill Clintonôs motherôs maiden nameò. The secondary, general purpose browsing task was: ñfind all the information on the relationship between carbon monoxide and desertificationò Following their completion of the task (where a search engine was definitely used), participants had their search actions played back to them and they were asked why they engaged in the actions that they did whilst performing the task. Notes of their responses were taken by the researcher during this process. To ensure that all the participants had the same experience, they all started the task from the same website (Wits Universityôs homepage). After each participant had completed the task the researcher cleared all browsing data (browsing history, download history, emptying the cache, etc) to make sure the following participant could not follow the search path used by anybody before them.

After participants had finished providing an explanation of the actions they performed while completing the directed search task, they completed a 3 item confidence and satisfaction 5-point Likert type scale (see Appendix 6) constructed by the researcher. Participants were then asked to provide an illustration of how they thought a search engine works accompanied by a brief written description to complement their illustrations should they have been difficult to interpret accurately by the researcher. Participation in this research (from the researcher briefing participants through to the drawing of how a search engine works) ranged from 15 to 55 minutes.

2.5 Measures

The following are the measures that were used in the study to obtain data.

Retrospective protocol interview itinerary: This interview itinerary was a list of questions that addressed the recorded search behaviours of participantsô performing the directed search task. The questions enquired about the various reasons participants opted for particular search actions during the process, in order to help determine the kind of understanding they had of how a search engine works (see appendix 8). Some of the questions asked were ñwhy did you click on that linkò, ñwhy did you choose that resultò and ñwhy did you use those search termsò.

The biographical questionnaire: Age, highest level of education and the institution it was obtained from, gender, race, language and occupational status were recorded using this questionnaire (see appendix 4).

A diagram of how users think a search engine works: This measure allowed for the effective assessment of how participants thought a search engine works. It is inspired by and similar to the diagrammatic impression Kerr (1990) asked users to provide of how they thought a database worked; and identical to the diagrammatic format Thatcher and Greyling (2003) used. This measure was used to help capture the mental models of users concerning the nature of search engines (see appendix 5).

Composite WWW experience: The average time participants spent on the WWW weekly; their preferred browser and the location/s used to surf the WWW; the kind of tasks they used the WWW for; their perceived level of competence using the WWW and description of how they learned how to use WWW were all aspects that were assessed using this questionnaire (Thatcher, 2003) (see appendix 3).Participantsô self-rated capability using the WWW was used as the measure for participantsô general WWW experience (i.e. composite WWW experience).

Performance measure: A mark sheet on which the researcher kept score of the *number of steps* taken to perform the directed search task, *time taken* to complete the directed search task, and the *accuracy* of the answers provided by the participants was used as the performance measure (see appendix 7). A step was defined as any action (e.g. entering a search term, going to a previous webpage, clicking on a link or opening a new tab) that participants undertook during the task/s. *Time* was defined as the recorded length of time that had elapsed from the beginning of the task up until participants indicated that they were done. Accuracy was defined as the extent to which the answer provided by participants matched the model answers (i.e. Cassidy or Blythe). As a result, participants were marked as either right or wrong.

Satisfaction and confidence measure: Satisfaction and confidence measures were taken on a 5-point Likert scale and constructed by the researcher. This 3-item measure provided information on the level of satisfaction experienced by participants regarding the whole search task and the manner in which they got their answers. It also measured the level of confidence that participants had in the answers that they provided (see appendix 6). Each item was measured individually.

2.6 Equipment

HyperCam 3 Software: This software provided the programme that made recording and playing back participants search actions possible. The software assisted in providing an accurate timer and thus negated the need for a stop watch. It also helped in retracing any anomalies in the data as the recordings were stored in avi format and played back as often as was necessary. More information on this software can be found at http://www.solveigmm.com/en/products/hypercam/

Desktop/Laptop: Both a desktop and laptop (hardware) with internet connection were used in order to allow participants to complete the task/s. Both the laptop and desktop were HP computers and they both used Microsoft Windows XP Professional; the 2002 version. The HyperCam 3 software used to aid in recording participants search actions was installed on both.

2.7 Analysis

As this research study was comparing performance and the quality of the mental models formed by the two samples from 2000 and 2010 , a few statistical techniques were employed. To assess whether differences existed regarding performance (i.e. steps and time taken at time one as compared to time two) on the directed search task, a two independent sample t-test was used (Huck, 2004). This test allowed any differences in performance that may have existed between the two samples to be determined and effectively evaluated (Huck, 2004). A chi-squared test was also used to assess the accuracy of the answer given by participants as this also formed part of the performance measure (Huck, 2004). The level of satisfaction with the whole search process was also analysed using the two independent sample t-test (Huck, 2004).

In order to assess whether differences existed between the mental models formed by users from the respective generations certain steps were taken. Seventeen salient features of search engines were identified in the original study (Thatcher & Greyling, 2003). Information provided by experts on search engines as well as trends identified in participants mental models of how search engines work helped in composing the seventeen salient features. These features were used by those researchers to perform a Wardȏs cluster analysis (Ward, 1963) that allowed them to determine possible mental model categories. The current study used that data as a point of reference. In order to determine the mental model clusters of the 2010 sample an identical (Wards) cluster analysis was conducted using the same salient features identified in the original study as no new features had emerged.

It was anticipated that the clusters of the 2010 sample would not be identical to those found in the 2000 sample. If that was not the case, a chi-squared test would be used to assess whether differences existed between the identical clusters (nominal variables) of the two samples (Huck, 2004). In the event that the clusters were not identical (i.e. clusters of respective samples did not meet the conditions for the stated statistical (tests) comparisons) a more descriptive analysis would be provided highlighting *what* those differences were and their *meaning* in the context of this research study. This descriptive analysis would also be based on chi-squared tests performed for comparisons between individual salient features for the respective samples (i.e. potential differences in salient feature 1 at time 1 and time 2; potential differences in salient feature 2 at time 1 and time 2; potential differences in salient feature 3 at time 1 and time 2, etc). As such, this description would *not* be in the form of any qualitative technique, but descriptive purely in the literal sense with the chi-squared technique forming the basis of the explication.

In their study, Thatcher and Greyling (2003) used two raters to determine the salient general features of search engines identified based on their participantsô drawings and written descriptions along with the mental model task (directed search task). Once these salient features were determined the two raters assigned them to each of the participants. A comparison was then conducted (Cohenôs Kappa) to ensure accurate ôfeature designationô to each of the participants occurred. The current study used the salient features identified in that study in coding the data obtained through the retrospective interviews and diagrammatic representations in a quantitative manner. This would be followed by the calculation of Cohenôs Kappa to ensure that the designations by the two raters used in the present study, were done accurately (Huck, 2004; Thatcher & Greyling, 2003).

2.8 Ethical considerations

All participants were provided with a participant information sheet prior to their participation. It was explained that their participation would be considered as consent on their part. While all participants had their search actions recorded through onscreen capture software and by the researcher as he took notes related to their search actions, there was no need to ask for written consent as participants themselves would not be recorded by video or any means whatsoever requiring written consent. As a number of the participants required for the study were high school students, a parental consent form was provided to ensure the necessary consent for their ethical participation was received. The researcher made it clear that participation in the study was on a voluntary basis, that participants could withdraw at any stage during the study without prejudice, and that their confidentiality and anonymity was guaranteed. Confidentiality of the participants was ensured through the researcherõs actions to keep identification hidden as codes were assigned to each participant to effectively link retrospective information to participant search recordings. A detailed description of the research was also provided to participants to help facilitate a better understanding of the study and their own contributions to it. Obviously, information related to the task itself was explained in an indirect way so the integrity of the results was not affected.

The directed search task and the general purpose browsing task were not of a provocative nature and were not emotionally charged at all (participants were asked to ñfind Bill Clintonõs motherõs maiden nameò for the directed search task and to ñfind all the information on the relationship between carbon monoxide and desertificationò for the general purpose browsing task). Therefore no adverse emotional and psychological effects were expected of anyone participating in the study. The researcher provided his contact details along with those of his supervisor for participants who wanted to have any queries or concerns addressed. A summary of the results was made available on a blog set up by the researcher.

Chapter 3

3.1 RESULTS

The following chapter will discuss the results obtained from the statistical analyses conducted. A description of the results found through the use of Cohenꞌs Kappa for the inter-rater reliability, Independent sample t-tests for performance regarding *time* and *number of steps*, Chi squared test for performance regarding the *accuracy* of the answer, and cluster analyses for determining the mental model clusters for the 2010 sample will be provided. As the clusters of the respective samples will be shown to have not been identical, results of the chi-squared tests on the salient features of search engines that showed differences will be presented. The results of the Independent sample t-tests for satisfaction will then be presented and further elaboration on all the results will be provided in the discussion chapter.

3.2 Inter-rater reliability

In order to ensure that the salient features of participantsꞌ mental models were accurately assigned two raters were used to carry out this task. Both raters assigned salient features to all 80 participants. When a Cohenꞌs Kappa was run to assess the level of agreement it was found that a weighted Kappa coefficient of 92 percent with 95 % confidence limits between lower 87 and 97 was present. The weighted Kappa coefficient of 92 % (there was 91% agreement in 2000) indicated that a high level of agreement was present between the two raters in the designation of salient features. Following an inspection of the eight inconsistencies that emerged between the two raters it was discovered that minor interpretive issues had caused confusion and consensus was quickly reached following their resolution.

3.3 Distribution analysis

In order to ensure that the correct statistical techniques were used a distribution analysis was conducted on all interval variables. This would assist in determining whether the variables matched parametric or non-parametric parameters and thus aid in guiding the most relevant method of approaching the analysis. As the only variables of a purely interval nature were *time* and *number of steps* the prime focus was on those two. The Satisfaction and Confidence scale only had 3 items that were measured separately and thus treated individually. The distribution analyses results for *time* and *number of steps* revealed that they were both normally distributed. The results for these two variables are summarised in *table 1* below.

Table 3: Distribution analysis of interval variables

Variable	**Mean**	**Median**	**Histogram**	**Distribution**	**Parameters**
Time	359.668	269.50	Normal	Normal	Parametric
Number of steps	11.36	9.50	Normal	Normal	Parametric
Satisfaction with whole search process	3.63	4.00	Non-normal	Non-normal	Non-parametric
Satisfaction with the method used to get to the answer	3.41	4.00	Non-normal	Non-normal	Non-parametric
Confidence with the answer	3.81	4.00	Non-normal	Non-normal	Non-parametric

3.4 Differences in user-performance on the directed search task over time; regarding steps taken, time, and accuracy (research question 1)

Two independent sample t-tests were used to assess whether differences existed in the performance of the two respective samples (2000 and 2010) regarding both *the number of steps* and *time* taken to task completion. Statistical significance was found for the difference in the number of steps taken to task completion ($t\ (134.07) = 5.59$, $p < .0001$) with the 2010 sample taking significantly fewer steps to complete the directed search task. Both the pooled and satterthwaite methods indicated that a statistically significant difference exists. Given that the variances between the two samples were statistically unequal the satterthwaite method was used as the interpretive tool. Please see *table 4* for a summary of these results.

Table 4: Two independent sample t-test results for steps and time

	Mean 2000 (SD)	Mean 2010 (SD)	t-value
Steps	14.25 (7.80)	8.48 (4.97)	5.59 **
Time	497.10 (272.6)	222.20 (155.9)	7.83**

** $p < 0.01$

Statistical significance was found for the difference in the time taken to task completion ($t\ (125.65) = 7.83$, $p < .0001$) with the 2010 sample taking significantly less time to complete the directed search task. Sattherwaite was again used as the interpretive tool as the variances of the respective samples were again statistically unequal (*see table 4*).

The difference in performance for the respective samples was so great for both *time* taken and *number of steps* that an analysis of covariance (ANCOVA) was conducted to identify any possible covariates that might have affected performance. The variable that was identified as the most likely covariate, following t-test comparisons that revealed statistical significance (see table 5), was participantsô self-rated web experience capability. After covarying for participantsô self-rated web experience for both *number of steps* ($F(2)= 15.60$, $p <.0001$) and *time* ($F(2)= 32.50$, $p <.0001$) the statistically significant differences between the two samples were still apparent.

Table 5: Independent sample t-test results for self-rated web experience

	Mean 2000 (SD)	Mean 2010 (SD)	t-value
Self-rated WWW experience	3.32 (0.89)	3.70 (0.75)	-2.86**

** $p < 0.05$

There was a clear indication that a statistically significant difference in performance, regarding both time and number of steps, existed even after the covariate of self-rated (composite web experience) was factored in.

A chi-squared test conducted to assess whether there were differences in performance (see *table 8*) regarding the accuracy of the answer provided by the respective samples indicated that there was a statistically significant difference ($G^2 = 4.68$, $p < 0.05$).

Table 6: Chi-squared result for accuracy of answers

Correct Answers at T1	Correct Answers at T2	G^2
46	59	4.68**

**$p < 0.05$

The results for performance therefore indicated that there was a clear difference in performance measures *time* and *number of steps*, with the 2010 sample requiring significantly less time and fewer steps to complete the task. The accuracy of the answer provided by the respective samples indicated that a statistically significant difference does exist, with the 2010 sample providing significantly more correct answers. This outcome will be explored more in the discussion section in the following chapter.

3.5 Salient features

Seventeen salient features divided into four categories (i.e. choosing a search engine, formulating a query, executing the query and examining search engine results) were used to help determine the accuracy and completeness of participantsô mental models of search engines. The table below (see table 7) provides a summation of what these salient features were so the analyses described after this are effectively contextualised.

Table 7: Salient features used

Salient Feature Category	Salient Features
Choosing a search engine	1: Chooses a search engine rather than letting the web browser decide (SF1)
	2: Recognises that a search engine is a database (SF2)
	3: No understanding of the web as the collection place for a database (SF3)
	4: Understands that search engines use a specific algorithm to collect the database (SF4)
	5: Recognises that different search engines have different databases (SF5)
Formulating the query	6: Participants choose keywords or phrases to search (SF6)
	7: Participants browse search engine categories to widen/narrow the search domain (SF7)
	8: Participants use Boolean logic or other operators to change search parameters (SF8)

	9: Understanding that search terms can be changed/modified to widen/narrow the search of the Web or database (SF9)
Executing query	10: Search engine matches terms/phrases to the identifiers on the webpage or database (SF10)
	11: Search engine matches terms/phrases to words in webpage or Web document (SF11)
	12: Search engine matches terms/phrases with the webpage title (SF12)
	13: Search engine looks for terms/phrases and related terms/phrases or extensions (SF13)
Examining search engine results	14: Search engine algorithm ranks results (SF14)
	15: Different search engines have different ranking algorithms (SF15)
	16: Search engines allow users to narrow search results by finding ñsimilarò results (other relevant/connected results) (SF16)
	17: Search engine displays a hyperlink to the original location of the information (SF17)

3.6 Mental model clusters

The cluster analysis results from the 2000 sample will be briefly discussed first to aid in a better understanding of the current studyôs cluster analysis results, which will be provided after. A table is also provided to show the frequency of the salient features at T1 and T2 to provide further comparative clarity (see table 8).

Table 8: Frequency of salient features at T1 and T2

Salient features	Frequency at T1 2000	Frequency at T2 2010
SF1	36	74
SF2	10	17
SF3	26	61
SF4	8	12
SF5	73	9
SF6	22	73
SF7	11	3
SF8	52	16
SF9	6	39
SF10	4	41
SF11	54	2
SF12	12	1
SF13	0	0
SF14	4	30
SF15	0	7
SF16	18	1
SF17	22	1

Cluster analyses including the Pseudo T^2 statistic, cubic clustering criterion (CCC) and the dendogram were conducted for the sample from 2000. The CCC showed slope changes at 4/5 clusters while the dendogram indicated two very distinct clusters which both divided into two more clusters. Following careful examination of all the results, it was resolved that four clusters provided the most meaningful and concise description of the data.

The four distinct clusters were reasonably similar in size as they ranged from N=16 for cluster 2 to N=26 for cluster 4. Cluster 1 had 17 participants and cluster 3 had 21 participants.

3.6.1 2000 Cluster 1 (N=17)

This cluster, along with cluster 2, was reported as having the simplest mental models and recognised very few salient features. Participants in this cluster were characterised by a distinct lack of awareness that; multiple search engines exist; search engines search their own databases and not the entire WWW; and that search engine results are ranked. Participants in cluster 1 were distinct in that they identified keywords as being important in the quality of the search process and the consequent matching of those keywords to the information retrieved.

3.6.2 2000 Cluster 2 (N=16)

This cluster also had very simple mental models and had an identical lack of understanding to that of cluster 1 regarding salient features. That is, Cluster 2 shared the same qualities as cluster 1. Cluster 2 participants were distinct in that they placed greater emphasis on searching for information using browser-defined categories. They did not identify many more features.

3.6.3 2000 Cluster 3 (N=21)

Cluster 3 participants, similarly to those in the fourth cluster, showed more detailed mental models, identifying more salient features than the first two clusters. Concepts that were nebulous to the first two clusters such as recognition of multiple search engines, search engines locating information within their databases as obtained from the WWW, and the understanding that refining search terms influences the quality of the search outcome were present in this cluster. Most notably, all the participants in this cluster showed a clear awareness of multiple search engines and close to half of these participants indicated that different search engines used different algorithms to get their results.

3.6.4 2000 Cluster 4 (N=26)

Participants in this cluster showed mental models that were almost identical to those in cluster 3. That is, they too showed a clear understanding of the presence of multiple search engines; the role refining a search has on the outcome, and search engines as databases. This cluster was distinct in that many more participants showed an understanding that search engines derive their databases from the WWW and significantly fewer participants showed an understanding of different search engines having different algorithms for producing different results.

The cluster analysis results for the 2010 sample indicated that there were three distinct clusters. Each cluster was characterised by certain salient features of search engines and allowed for a sensible classification, by individual cluster and generally, of the kind of mental models that were present in the sample. The clusters were of reasonably similar sizes ranging from N=18 in cluster 3 to N= 34 in cluster 2, with cluster 1 containing 28 participants. In all of the clusters participants tended to choose their own search engine instead of letting the Web browser ǵdecideǭ This was a similar feature that was common across the entire 2010 sample. Another common feature across all three clusters was that participants used key words and phrases to search. All three clusters showed little evidence of understanding that the use of Boolean logic could help refine a search. Indeed, not a single participant identified this in the first cluster and 64% of participants in cluster 2 and 77% of participants in cluster 3 did not identify this. Participants in Cluster 3 showed a complete understanding of a search engine being a collection of databases whereas clusters 1 and 2 did not.

In all three clusters participants showed an acceptable understanding of the impact modifying search terms can have on the scope of the search itself with over 55% of participants in clusters 2 and 3, and just under 40% of participants in cluster1 indicating this understanding. None of the participants in any of the clusters showed an understanding that search engines display a hyperlink to the original location of the information. Just fewer than 50% of participants in clusters 2 and 3, and 30 % of participants in cluster 1 showed an understanding that different search engine algorithms rank search results; however, only 28% of participants in cluster 3 identified that different search engines have different ranking algorithms. Less than 4% of participants in clusters 1 and 2 showed this understanding.

Table 9: Percentages of the salient features across the 2010 clusters

Salient Feature	2010 Cluster 1 percentage	2010 Cluster 2 percentage	2010 Cluster 3 percentage
1	88.88%	94.11%	94.44%
2	0%	0%	94.44%
3	100%	97.05%	0%
4	3.70%	2.94%	55.55%
5	3.70%	1.47%	16.66%
6	96.29%	85.29%	94.44%
7	3.70%	2.94%	5.55%
8	0%	35.29%	22.22%
9	33.33%	58.82%	55.55%
10	0%	94.11%	50%
11	0%	2.94%	5.55%
12	0%	2.94%	0%
13	0%	0%	0%
14	44.44%	2.94%	44.44%
15	3.70%	2.94%	27.77%
16	3.70%	0%	0%
17	0%	0%	5.55%

Individually, the characteristics of the three 2010 clusters indicated significantly different levels of understanding regarding the mechanics of search engines.

2010 Cluster analysis

3.6.5 2010 Cluster 1 (N=27)

Participants in this cluster showed a distinctly poor understanding of how search engines work with very little understanding of many salient features of search engines. The most distinctive features of this cluster were that participants placed great emphasis on using key words in their search; they chose their own web browser, and had no idea that a search engine is a database (i.e. no participants in this cluster identified this). Close to 45% of participants in this cluster showed some understanding of search results being ranked. The mental models of participants in this cluster were thus very simple.

3.6.6 2010 Cluster 2 (N=34)

Participants in this cluster showed a slightly more advanced level of understanding regarding search engines. A very strong understanding of search engines matching search terms and phrases to identifiers on a webpage or database was present in this cluster. There was also a moderately appreciable increase in understanding regarding the role of Boolean logic and a strong understanding of the impact changing/modifying search terms has on widening or narrowing the scope of the search. Similarly to the cluster 1 though, almost all participants in this cluster were unaware that a search engine is a database.

3.6.7 2010 Cluster 3(N=18)

This cluster showed the most advanced level of understanding with a relatively strong indication of knowledge regarding the fact that a search engine is a database, that search engines use specific algorithms to collect their database, that key words and search modifications have an impact on the quality of the search, and that search engine algorithms rank results. No participants in this cluster failed to identify a search engine as a database. A strong appreciation for the importance of keywords or phrases in a search was also apparent for this cluster.

3.7 Cluster analysis summary

The 3 clusters of the 2010 sample shared a few similarities with the clusters found in the 2000 sample. Like in the 2000 clusters, a clear indication of an increment in the quality of mental models across the 2010 clusters was present. The clusters (1 and 2 in 2000 and 1 in 2010) with the simplest mental models in both samples showed little understanding of most of the salient features, with a poor understanding of a search engine as a database and the existence of multiple search engines. However, unlike clusters 1 and 2 in the 2000 sample, where there were no participants who showed an understanding of search engine results being ranked, close to 50% of cluster 1 participants in the 2010 sample did. That is, even across the clusters with the simplest mental models between the two samples, there was a reasonable indication that the 2010 sample had ñless simpleò mental models. Indeed, the fact that there was only one cluster that emerged as being undeniably basic in 2010 as opposed to the two that emerged in the 2000 cluster was probably indication enough that improvements in mental models between the two sampleôs clusters were present. Cluster 2 of the 2010 sample, which was slightly more advanced than cluster 1, indicates a substantial improvement regarding the grasp of salient features. While the difference was slight with cluster 1, cluster 2 of the 2010 sample represents the extent of the gap between the two samplesô simpler mental models with a greater appreciation for Boolean logic and knowledge of search terms matching a webpage or database over and above some of the qualities shared with cluster 1 of the 2010 sample.

Regarding the more advanced clusters of both samples, a clear appreciation of search engines as databases that collect information from the web was apparent. Also, the importance of keywords and knowledge that search engines rank results was present in both samples. While cluster 3 of the 2000 sample emphasised the fact that different search engines produced different results and cluster 4 of the same sample emphasised the importance of keywords matching with databases, these qualities were both emphasised in cluster 3 of the 2010 sample along with the importance of changing/modifying search terms to influence the quality of the search outcome. Recognition of multiple search engines for the advanced cluster (3) of 2010, while present, was not as strong as it was for clusters 3 and 4 in the 2000 sample. In that regard, the óadvancedô clusters of the respective samples, even with some similarities when assessed in an all encompassing manner, were clearly quite different when separated and analysed accordingly (i.e. cluster 3 of 2000 and cluster 3 of 2010 and cluster 4 of 2000 with cluster 3 of 2010).

As such, while the clusters of 2000 sample were classified asô ósimpleô or óadvancedô, the clusters of the 2010 sample were ósimpleô, óslightly more advancedô and óadvancedô regarding mental models. Thus, the results indicated an appreciable difference between the two samples. This was an anticipated outcome and the intention to perform a chi-square to compare the clusters from the respective samples was not possible because there was no basis for comparison. As such, results from the chi-square performed on each of the salient features for the two samples will be reported.

3.8 Salient feature comparisons

There were seventeen salient features assessed for possible differences between the two samples using the chi-squared technique. The tables below summarise the range of the total number of salient features obtained by all participants in each sample and, more importantly, those that *showed* statistical significance.

Table 10: Comparison of the range and mean number of the total salient features at T1 and T2

Sample	Least no of salient features	Most number of salient features	Mean number of salient features
T1	0	10	4.20
T2	2	9	4.83

Table 11: *Chi-squared results for significant (more frequent) salient features between samples*

(note: non-significant results to be addressed in the discussion)

Greater frequency of:	Frequency at T1	Frequency at T2	Chi-square statistic G^2
SF 1at T2	36	74	42.00**
SF 3 at T2	26	61	30.86**
SF 5 at T1	73	9	102.46**
SF 6 at T2	22	73	67.39**
SF 7 at T1	11	3	5.00*
SF 8 at T1	52	16	33.14**
SF 9 at T2	6	39	33.66**
SF10 at T2	4	41	42.32**
SF 11 at T1	54	2	74.28**
SF 12 at T1	12	1	10.13**
SF 14 at T2	4	30	25.24**
SF 15 at T2	0	7	7.32**
SF 16 at T1	18	1	17.26**
SF 17 at T1	22	1	22.39**

* $p < 0.05$; ** $p < 0.01$

A statistically significant difference ($G^2 = 42.00$, $p < 0.0001$) was found for the first salient feature. Significantly more participants in the 2010 sample chose a search engine instead of allowing the Web browser to decide. A statistically significant difference ($G^2 = 30.86$, $p < 0.0001$)

was found for the third salient feature with significantly more participants in the 2010 sample showing no understanding of the web as the collection for a database.

There was a statistically significant difference ($G^2 = 102.46, p < 0.0001$) for the fifth salient feature. A significant number of participants in the 2000 sample showed an understanding that different search engines have different database algorithms. A statistically significant difference ($G^2 = 67.39, p < 0.0001$) was found for the sixth salient feature and this was indicative in that more participants in the 2010 sample used key words or phrases to perform the search task. A statistically significant difference ($G^2 = 5.00, p < 0.05$) was found for the seventh salient feature. More participants from 2000 (T1) browsed search engine categories to widen/narrow the search domain There was a statistically significant difference ($G^2 = 33.14, p < 0.0001$) for the eighth salient feature, with more participants at T1 using Boolean logic and other operators to change their search parameters. A statistically significant difference ($G^2 = 33.66, p < 0.0001$) was found for the ninth salient feature. A greater number of participants in the 2010 sample showed an appreciation for the impact changing search terms can have on modifying (widening or narrowing) the search of the web or database. A statistically significant difference ($G^2 = 42.32, p < 0.0001$) was present for the tenth salient feature. A significantly larger number of participants in the 2010 sample showed an understanding that search engines match search terms and phrases to the identifiers on a webpage or database. A significant difference ($G^2 = 74.28, p < 0.0001$) was found for the eleventh feature as well. Significantly more participants from time 1 understood that search engines match terms and phrases to in the actual webpage or web document. There was a statistically significant difference ($G^2 = 10.13, p < 0.05$) for salient feature number twelve. More participants from time 1 understood that search engines match terms and phrases with the webpage title. There was a statistically significant difference ($G^2 = 25.24, p < 0.0001$) found for the 14^{th} salient feature with many more participants in the 2010 sample showing an understanding that search engine algorithms rank results. A statistically significant difference ($G^2 = 7.32, p < 0.05$) was found for this salient feature. More participants from the T2 sample showed an understanding that different search engines have different ranking algorithms even if this number was, technically speaking, quite small. A statistically significant difference ($G^2 = 17.26, p < 0.0001$) was found for salient feature 16 as well with significantly more participants from the 2000 sample indicating an understanding that search engines allow users to narrow search results by finding ñsimilarò results (other relevant/connected results).There was a statistically significant difference ($G^2 = 22.39 p < 0.0001$) present for the 17th salient feature. More participants from time 1 understood that search engines display a hyperlink to the original location of the information. Broadly speaking, the number of participants who knew this was still paltry, but statistically the difference was very significant.

<u>3.9 Comparison of levels of satisfaction regarding the whole search process</u> <u>*(Revert to table 3 for scores of the other two items for 2010 (note: no measures of confidence (item 3) and satisfaction regarding how the answer was reached (item 2) were taken in 2000)*</u>

No statistical significance ($p > .05$) was found for the level of satisfaction regarding the whole search between the two samples (item 1). This indicated that both samples had similar levels of satisfaction. These scores indicated feelings of neutrality across the two samples, with a slight indication that the 2010 sample was moving towards being very satisfied with their whole search experience.

<u>*Table 12: Two independent sample t-test results for satisfaction with the whole search process*</u>

	Mean 2000 (SD)	Mean 2010 (SD)	t-value
Satisfaction with the whole search process	3.49	3.63	-0.78

* p < 0.05

<u>*Table 13: Comparison of search engines used at T1 and T2 (note: 2 participants in each sample did not use a search engine for the directed search task)*</u>

Search engine chosen	Number of participants at T1	Number of participants at T2
AltaVista	9	0
Google	3	69
Yahoo	29	1

Bing	0	8
Looksmart	9	0
Ask(jeeves)	8	0
Infoseek	9	0
Metacrawler (metasearch engine)	1	0
Lycos	2	0
Hotbot	1	0
SNAP	1	0
Excite	1	0
Inforocket	1	0
Dogpile (metasearch engine)	2	0
Search.com	2	0

The results indicate that a larger number of different search engines (and metasearch engines) were used by participants at T1. However, the majority of participants at T2 used Google.

3.10 Results summary

Significant differences were found for all the performance measures, with the 2010 sample requiring significantly fewer steps and less time to complete the directed search task. The 2010 sample was also significantly more *accurate* in the answers they provided. The cluster analysis for the 2010 sample revealed three distinct clusters in contrast to the four that were present in 2000. The clusters of the respective samples shared a few similarities but were, in the main, quite different. Consequently, chi-squared comparisons of the salient features were performed and indicated that seven of the seventeen features were statistically more frequent at Time 2 and seven of the salient features were statistically more frequent at Time 1. There was no statistical difference in satisfaction regarding the whole search process between the two samples.

Chapter 4

4.1 DISCUSSION AND CONCLUSION

In the following chapter, all of the reported results will be discussed in order to determine what they actually mean. The results will also be discussed in the context of past research to establish whether similar trends exist or whether new patterns have emerged. Discussing these results will allow for meaningful inferences to be made regarding the practical implications of the findings for both future research and certain design aspects of search engines. The findings will be discussed sequentially but not in isolation. This is because a number of the findings are linked and discussing them in conjunction provides a more holistic understanding.

4.2 Differences in user-performance on the directed search task over time; specifically regarding steps, time and accuracy (research question 1)

Significantly fewer steps were required by the 2010 sample as opposed to the 2000 sample in completing the directed search task. Borgman (1986) and other authors (e.g. Staggers & Norcio, 1993; Doyle & Ford 1999) have argued that the more complete and accurate usersô mental models of a particular system become, the better users interact with that system. This is possibly one of the main reasons why the 2010 sample showed greater efficiency in the number of steps used. Craik (1943) argued that more accurate and complete mental models aid our understanding of all the systems we engage with (e.g. computers, the natural environment, etc) and noted that a likely consequence of this was a more parsimonious approach to interactions with those relevant systems as mental models developed. There are other possible influencers in this regard though. The databases of search engines have grown substantially in the last ten years and information retrieval has subsequently become less tedious. Sullivan (2006) notes that as a result of substantially larger information databases information triangulation has become a lot faster for searches. Web crawlers and/or spiders (i.e. the programmes that search engines use to retrieve information from the WWW for their databases) now operate on a principle of órelevanceô and deliberately seek out information that is most popular amongst users (i.e. what is most searched for) or information that is related (in any number of different ways) to the information that is most popular to users (Sullivan, 2006). This órelevanceô system has meant that data for a massive range of popular topics is not only abundant, but is available from many different perspectives. This aspect was apparent in that the information sought from the directed search task was retrieved efficiently even when usersô queries were conceptually very different. For instance, those participants who searched óBill Clinton family treeô (roughly 10% of the sample), those (roughly 40%) who searched ñBill Clintonôs motherò and those (roughly 7%) who searched ñBill Clinton family historyò, mostly used more than 4 steps and less than 10, an observation consistent with the results of this particular performance measure. This is probably evidence of how quickly information converges because of the multiple links established, even when the initial searches were quite different. Even participants with seemingly less concise/refined searches (e.g. ñhistory of all U.S/American presidentsò or ñWho did Bill Clintonôs father marryò, etc) required few steps. In fact, those participants who required more than 10 steps were generally those who wanted to cross-check their first answer (also retrieved with relatively fewer steps) against other sources. Thus, the comprehensive nature of search engine databases is also a possible reason why significantly fewer steps were required. It is also possible that new functional aspects of search engines such as ñpredictive drop-down menusò (a consequent feature of the órelevanceô or ópopularityô principle search engine crawlers and spiders use) that give possible options as users type in their search query could have played a role in the significantly fewer step-laden process for the 2010 sample. Spink et al. (2000; 2001) and Muramatsu and Pratt (2001) noted that usersô query formulation was not particularly good and thus did not facilitate an efficient method of retrieving information. Given the functional modifications that search engines have undergone in the last 10 years, it is plausible that these modifications played a role in facilitating a more efficient, step-reduced process. Undoubtedly, and this is a theme throughout the discussion, the substantial increase in the frequency of use of the internet, and by implication search engines, means that users in the 2010 sample, over and above having mechanisms that aided them in their query formulation were, in the main, quite familiar with the process of query formulation and that may have also contributed to significantly fewer steps being used by the 2010 sample.

That is, the significant difference found for self-rated web experience between the two samples, where the 2010 sample showed significantly greater self-rated web experience, probably had an influence regarding the number of steps used and time taken as well. The measure used to encapsulate composite web experience was *self-rated web experience* and a significant difference was found. This measure was seen as the best summation of; how long participants had been using the WWW; how often they used the WWW; how many hours a week they used the web; and what they were using the WWW for. Unsurprisingly, without exception, the 2010 sample showed significantly greater differences in all these aspects (see table 2 for summation). Participants from the 2010 sample had, on average, over 80 months more web exposure and spent more than double the time, per week, on the internet. They also used the WWW more frequently. All this information makes it all the less surprising that a significant difference in self-rated web experience (composite web experience) was present. The consequences of the individual aspects of composite web experience do, however, require discussion in the context of the possibility that usersô in the 2010 sample formulated more complete and accurate mental models.

Theorists such as Borgman (1986), Staggers and Norcio (1993) and Slone (2002) all highlight consistent system exposure, directly and/or indirectly, as the key to forming more complete and accurate mental models. Craik (1943) also described the quality of oneҫs *understanding* as being a function of experience (of course room must be made for individual differences here). Given that both the length of use of the WWW and number of hours a week spent on the WWW were far greater for the 2010 sample, it is evident that opportunity for forming more complete and accurate mental models was present. In fact, the various queries and search methods employed by the 2010 sample were consistently creative, with a number of participants commenting that they had had experiences with particular websites, search styles and search engines that influenced how they searched. Consistent with the theory and results reported by Muramatsu and Pratt (2001) and Slone (2002) amongst others, it makes sense that the 2010 sample showed stronger signs of more complete and accurate mental models given the differences in exposure and the consequent significant difference in composite web experience. However, the role of the advancements made in the technological applications of devices like cell phones and 3G to the WWW, which a number of participants indicated using, must be considered as an important facilitator in the level of exposure enjoyed by the 2010 sample that was not available a decade ago. It is a clear consequence of the ex post-facto variable time and is important to note this because it provides perspective on the differences in performance.

Significantly less time was required by the 2010 sample in completing the directed search task. Indeed, the 2010 sample took less than half the time in completing the task. There may be a number of reasons why this is the case. As noted above, the mental models of the 2010 sample, which are probably more complete and accurate than of the 2000 sample, might have assisted in facilitating task interactions that were much faster as a logical consequence of the fewer steps required. The tendency of quite a larger number of participants in the 2010 sample to search using techniques like opening new tabs and running multiple searches simultaneously, was evidence of a quite advanced level of understanding that stemmed from considerable experience (Thatcher, 2008). This evidence of more proficiency in using the system may have also played a role in increasing the 2010 sampleǿs speeds. Of course, and probably more so for this particular performance measure, there were other reasons why the gap in performance speed of the respective samples was so large. Firstly, the speed of the internet has increased quite dramatically since 2000 and the bandwidth now offered, in combination with the processing power of computers (e.g. dual cores, quad cores etc) lends itself to a considerably faster trial and error process that participants from the 2010 sample could go through. On a simple level, this means that if participants in the 2010 sample engaged in the exact same actions in reaching the desired answer as participants in 2000, they would still be considerably faster because of the difference in processing speeds. This is another consequence of the ex post-facto independent variable *time*. Spink et al. (2006), in analysing log file data for query formulation, noticed how little time users were spending on the internet irrespective of poor query formulations. They conjectured as to why this was the case and one of the answers reached was the speed of the internet. Spink et al. (2006) argued that increased internet speeds could have meant that even in the presence of poor query formulation and re-formulation by users, the information sought was retrieved relatively quickly. The speed of the internet has increased even more since 2000 and played a role in the significantly reduced time used by the 2010 sample.

The new features search engines have (e.g. predictive search text and dropdown search menus) also played a role in the significantly less time required by the 2010 sample to complete the directed search task. The new functional aspects of search engines and their significantly larger databases not only impacted *steps* but also *time*. They facilitated a faster search process and therefore reduced the time most participants spent completing the task (Spink et al., 2000). The fact that there are a lot more websites to retrieve information from now could have also played a role in the significantly faster time the 2010 sample took in completing the directed search task.

The 2010 sample had significantly more accurate answers in comparison to the 2000 sample. The reasons for this result are, in many ways, similar to the reasons for those elucidated for *time* and *steps*. The comprehensive databases search engines have, aided by alternative information databases (e.g. Wikipedia and Yahoo Answers) available to users\ played a role in significantly more accurate answers being provided by the 2010 sample. Also, and this was indicated by the methods some participants used to ensure their answers were correct (i.e. the use of Wikipedia and other search engines and information databases); the significantly greater frequency with which participants in 2010 used the internet (not necessarily their composite web experience *see table 2*) had also played a role as most 2010 participants mentioned how they often confirmed the accuracy of information retrieved using these and various other databases whenever they used the search engines. While the ex post-facto variable of time might have had a substantial impact on the performance measures *steps* and *time* (regarding speed and opportunities to interact with the system), it might have played a less substantial role for accuracy. Indeed, the fact that significantly more participants in 2010 sample got the correct answer is an indication that mental models of search engines have probably improved in the last decade. This is likely in that, unlike the performance measures *speed* and *time*, accuracy was more a function of the participantsô abilities to sift through and make sense of information. And while there are many helpful tools available to assist in validating data, there are equally, and perhaps more hindering aspects when using the internet. As such, the fact that a significant improvement was apparent in the retrieval of the desired answer by the 2010 sample in what Staggers and Norcio (2003) described as highly complex domain could be seen a clear marker of mental models of search engines having improved in the last ten years.

4.3 Differences in satisfaction regarding the whole search process

There was no significant difference in satisfaction regarding the whole search experience between the two samples in completing the directed search task. Most participants indicated neutrality about the experience. This is unsurprising as very few research participants participate out of a genuine interest in a particular study. Most do so because there might be an ethically sound reward or out of sympathy or reasons along that line. As such, when asked to rate their experience, they tend to be neither excited or offended and rate their experiences accordingly (Whitley, 2002)

<u>4.4 Usersô development of more accurate and complete mental models of search engines over time (research question 2)</u>

The clusters of the respective samples were found to be quite different to one another. This is perhaps unsurprising as the differences reported for all the performance measures gave some indication that this might be the case. Firstly, there were four clusters in the 2000 sample and three in the 2010 sample. Where the first two clusters represented simpler mental models of search engines and the third and fourth clusters represented more óadvancedô mental models in the 2000 sample, the 2010 sample had one ósimpleô cluster, one slightly more óadvancedô cluster and the óadvancedô mental models cluster. There were five more people in the more advanced clusters in 2010 compared to 2000. The more distinct characteristic of the 2010 sampleøs clusters breakdown might be a function of the large degrees of difference in the length of web use and the likely more pronounced variances that this might allow for. In terms of the comparison between the clusters, the simpler mental model clusters of both samples (clusters 1 and 2 for 2000 and cluster 1 for 2010) shared a few characteristics. For instance, in both of these ósimpleô clusters the failure to recognise multiple search engines and that a search engine is a database was apparent. However, unlike the ósimpleô clusters in 2000, close to half of the participants in the 2010 sampleøs simple cluster indicated some understanding of search results being ranked using algorithms. It is possible that the reason why the two simpler clusters share the same lack of understanding regarding search engines as databases and the existence of multiple search engines is because they represent the least experienced part of the respective samples. That is, people who have not used the WWW for a long period of time and are thus, logically, not likely to discover this information. As these participants have had limited access to the WWW, they might have not had the appropriate opportunity to explore the WWW. Equally, the ability of almost half of the participants in the 2010 sampleøs simpler cluster to note that search engines use algorithms to rate results might also be a function of the length of use of the WWW. That is, given the considerably large difference reported in length of WWW usage between the two samples, even the less ‘experienced’ participants in 2010 might have spent considerably more time using search engines and as such, had a greater opportunity to identify this particular feature of search engines. Another explanation, which is consistent with the observation made by Sullivan (2006) regarding the quality of information retrieved, is that information presented in search engine results pages is in a manner that highlights quite clearly that data is ranked from the most to the least relevant. This is a feature of search engines that was picked up by even the simplest cluster in 2010. The slightly more advanced cluster of 2010 showed an understanding of search engines matching search terms or phrases to web pages and databases as well as the impact of modifying search terms and using Boolean logic had on widening or narrowing a search. For these reasons this was a ‘standalone’ cluster in that it showed a slight improvement from the first 2010 cluster and was non-comparable with any of the 2000 clusters as they did not have convincing evidence of any of these features. There are a few possible reasons for this. Firstly, and this again reverts to the issue of WWW exposure, the extent of the difference in exposure between the samples hints at the possibility of a group of individuals in the 2010 cluster who represent a level of exposure that would not be found in the 2000 cluster. Put differently, participants in the 2000 cluster probably either had limited (simple clusters) or a lot of (advanced clusters) time on the WWW. As there are participants in the 2010 cluster who would have had more time on the WWW than the novice users and less time than participants with a lot of exposure, it is sensible that a cluster that represents such a group would emerge.

The more advanced clusters of the respective samples also shared similarities. They all identified the existence of multiple search engines to a fair degree, along with the fact that a search engine is a database and that refining searches influences the quality of the outcome. The differences lay in the fact that when compared individually (i.e. cluster 3 of 2000 with cluster 3 of 2010 and then cluster 4 of 2000 with cluster 3 of 2010) it was apparent that cluster 3 of the 2010 sample was substantially more advanced. That is, cluster 3 of the 2010 sample had most of the combined qualities of both cluster 3 and 4 in the 2000 sample. Cluster 3 of the 2010 sample also showed an understanding of the impact modifying search terms can have on the outcome. This difference can again be seen as being the result of experience. There is a clear indication that there were participants in the 2010 sample that had the same combined knowledge of the two ‘advanced’ clusters of 2000 and more. This represents a unique cluster that is also indicative of the more accurate and complete mental models present in the 2010 sample. This cluster probably represents individuals who have engaged the system frequently and would be consistent with the findings of researchers such as Muramatsu and Pratt (2001) and Doyle and Ford (1999) of more exposure improving mental models. It is important, however, to keep in mind that while the results suggest mental models are more complete and accurate a decade on, that this is only relative to the comparison being done. That is, even though more complete and accurate than the mental models of the 2000 participants, generally speaking mental models of search engines of the improved sample (i.e. 2010) remain still, in the main, quite poor with the strongest participants (2) identifying up to nine salient features.

4. 5 Salient features that are significantly increased in 2010

Significantly more participants in 2010 chose their own search engine in completing the directed search task. This result is indicative of the preferences of users of more recent times (i.e. the 2010 sample). This preference is probably a consequence of the greater use of the internet by the 2010 participants and the dominance of search engines such as Google in making themselves appealing to users. In becoming accustomed to different or particular search engines because of the considerably larger period of time they spend using the internet, 2010 users might have developed certain preferences (Liaw & Huang, 2004). Google was a popular choice amongst all 2010 participants. The default search engine Bing, was used by only 10% of participants with some even using it to get to Google. Liaw and Huang (2004) found that ñexperience with search engines significantly affects usersô attitudes toward search engines for information retrievalò (p. 501). This was evident in that when participants were asked about their choice of search engine, which was in the main Google, they generally pointed out that they had positive past experiences with Google or had heard great things about it. Indicative of this was that 95% of the participants in 2010 (see table 8 comparison of SF1) chose a search engine when compared to the 45% of the 2000 sample. They did not simply defer to the default Web browserôs search engine (i.e. Bing). In fact, a number of participants did not acknowledge Bing as a search engine at all. As there was no dominant search engine a decade ago, there might have been no basis for such specific preferences as evidenced in the differences between the two samples (Liaw & Huang, 2004; Thatcher & Greyling, 2003).

The 2010 sample showed a significantly poorer understanding of the web as the collection for search engine databases. That is, there was a significant failure by the 2010 sample in understanding that search engines used the Web to collect their information. A number of participants in the 2010 sample commented on how anything could be found on Google. This, along with the illustrations provided, seemed to imply that Google ñwasò the Web and not a database that used the Web to collect its information. This result can also be explained with reference to the non-coherent concept of cyber-space most people have (Graham & Metaxas, 2003). In fact, when asked to provide an illustration of how they thought a search engine works, most participants would begin by saying ñno ideaò. It is unsurprising that participants in 2000 were less inclined to believe that search engines were the Web and not a composition of information extracted from it because, unlike Google which gives the impression of searching the whole web, search engines operated on a relatively small and unrefined scope (Wang et al., 2000).

Significantly more participants from the 2010 sample used key words or phrases to search. This is not surprising as Muramatsu and Pratt (2001) found that query formulation improved with consistent exposure to search engines. The 2010 sample generally used the internet with greater frequency (refer to table 2) and thus showed a significantly better understanding of this salient feature. The search terms and phrases used by 2010 participants were concise, lateral and even remarkably obscure at times. The observed search behaviours indicated a general level of comfort with search engines. Even when aided by ñdrop-down menusò many participants elected to use their own search terms and phrases. They chose to ignore the ñdrop-downò feature that was not available ten years ago.

Significantly more search term modifications and changes were made by the 2010 sample in widening or narrowing their search. This is a result that would seem to contradict some of the results above as it would imply that 2010 participants were not getting the information they wanted. This is actually not the case. While some participants struggled, the majority of search changes were for confirmatory purposes as once participants felt they had found the correct answer they would type it in as a search and use Wikipedia and/or other knowledge databases to confirm the accuracy of their answers even though a number of them acknowledged that the information here provided was potentially incorrect (Wilkinson & Huberman, 2007). It is quite plausible that ten years ago changing search terms or modifying them was seen as almost meaningless and that more specific query formulation strategies such as the use of Boolean operators or other search engine categories was more helpful (Slone, 2002; Spink et al, 2001). Another possibility is that the slower internet speeds a decade ago made modifying search terms for confirmatory or any other purpose distinctly unappealing.

Participants in the 2010 sample showed a significantly greater awareness that search engines match terms/phrases to the identifiers on the webpage. This was evident in that participants consistently mentioned that they got a good indication of whether their queries were correct based on the string of highlighted terms which matched with their inputted searches. It is not immediately clear why the sample of 2000 did not recognise this. There are though, a number of factors that could provide an explanation. Results from the Thatcher and Greyling (2003) study indicated that users employed different search engines, and indiscriminately so. As such, they were probably unlikely to pick up that search engines used identifiers. That is, search engine set-ups are different (e.g. different algorithmic codes, functionally different spiders and crawlers, and different information interpretation tools etc) (Mowshowitz & Kawaguchi, 2005) and this means that their respective presentations of information also differ regarding what information is prioritised, how it is presented, and the accuracy of the information judged as important (Jansen & Spink, 2006). As such, the varying styles and levels of quality in the way search engines presented information might have confused users a decade ago.

Regarding the examination of search engine results, participants from the 2010 sample showed significantly more awareness of the fact that search engine algorithms rank results. Participants consistently remarked on how there was a definite ordering system and how it moved from the most relevant matches to the least relevant when asked why they opted for a result that they did. As most participants used Google, this was in fact a feature they said they found particularly useful as there was little need to skim through more than the first four search results. This is a feature that has clearly won over a great number of users (SEW, 2009) and is probably one of the reasons why users preferred Google to most other search engines. As previously stated, the search engine landscape ten years ago might not have been as clearly defined and as such, users might not have been able to pick up this particular feature as their trust was placed more in reading the web document and webpage titles as opposed to paying attention to the results provided.

4.6 Salient features that are significantly decreased in 2010

There was a significant difference regarding the recognition that different search engines have different database collection algorithms. The 2000 sample was more in tune with this salient feature indicating that, firstly, there was a willingness to try different search engines and secondly, that they were able to tell that the information across search engines was different. This difference is explained by the fact that in choosing their own search engine, most participants in the 2010 sample used Google (SEW, 2009). This would imply that they were not in a position to even make this observation because of their highly frequent use of this particular search engine. Sullivan (2006) observed the growing popularity of Google and the general perception held, corroborated by extensive market research (SEW, 2009), is that almost everyone everywhere uses Google now. Thatcher and Greyling (2003) noted the considerably less partisan support participants in the 2000 sample had regarding search engines. Perhaps the quality of information across search engines was perceived as being different enough to encourage the use of different search engines for different tasks ten years ago.

Participants from 2000 employed the tactic of browsing the search engine categories to widen/narrow the search domain significantly more than the 2010 sample. This result is consistent with the less sophisticated state of search engines in the past that encouraged users to use the full repertoire of search engine features (e.g. Entertainment, Science & Technology, etc) (Spink et al., 2001). Given the comprehensive databases search engines now have the need for use of more than the basic search has seemingly waned. Indeed, a number of participants, when quizzed on the other search engine categories, stated that there had never been a need to use them. Indeed, a large number of participants did not acknowledge the other categories at all. This is consistent with the fact that Google has done away with these categories.

Significantly greater use of Boolean logic and other operators was apparent in the 2000 sample. Their willingness to use this tool to narrow and widen their search is also probably indicative of the less comprehensive databases that search engines had that required more precise query formulation for adequate information retrieval (Slone, 2002; Spink et al, 2001). It is quite clear that the modifications that search engines have undergone have meant users now need to simply type in reasonably sensible queries that are without Boolean operators to get comprehensive results. Caffarella and Etizoini (2005) emphasised the need for search engines that were robust in processing natural language queries (queries without technical operators). They predicted that search engines were starting and would in fact continue to incorporate this feature as time goes on. This is evident in that some Boolean operators are built into Google for example. Caffarella and Etizoiniœs (2005) observation was accurate and clearly not applicable ten years ago. As the result would suggest, more technical search techniques (i.e. Boolean operators) may have been required to extract the relevant information in 2000 (Liaw & Huang, 2006; Slone, 2002).

A significantly greater understanding that search engines match terms and phrases with the web document was apparent in the 2000 sample. This result could be explained in that given the less sophisticated and perhaps precise presentation of webpage information a decade ago, a greater emphasis was placed by users on matches of search terms to the selected document. That is, confidence in the appropriateness of the search may not have been particularly high until a document was selected and opened so that a clear indication of the extent to which a match existed was determined. This is a result that also contributes to the understanding of the large speed differential between the two samples as this process would clearly have been more time consuming.

There were significantly greater levels of understanding that search engines match terms and phrases to the webpage title in the 2000 sample. Where participants in the 2010 sample were more than comfortable with using the highlighted extracts of the sub-text provided on the results page, this was not a real option for participants in 2000 as it was not available. The process of confirming the quality of information retrieved lay in assessing the extent to which typed searches matched with more ‗concrete' sources of retrieved data (i.e. webpage titles and web documents) as opposed to the summaries provided on the results page. This particular result probably highlights the varying levels of confidence participants from the two samples had regarding their ‗first-hit' search engine results, with participants from the T2 sample more than happy to scan for confirmation as opposed to assessing webpage titles and opening one web document after another to ensure the information matched the search (note: only after this was done would it be possible to actually assess the quality of the information retrieved). That was the process participants at T1 went through and also provides further explanation for the significantly greater number of steps participants in the T1 sample used.

There was a significantly greater awareness by the T1 sample that search engines allow users to narrow search results by finding similar results that are connected or relevant to the original set of results. This result is one that is expected. Given the less pronounced volume of information search engines had in their databases a decade ago, this particular feature would have been more noticeable then. Admittedly though, even though there was a significant difference here, less than a quarter of the sample in T1 was aware of this feature and much less still in the T2 sample. So while this result was significant it is important to note that both samples did not do particularly well in this regard.

4.7 Salient features that are not significantly different

There was no significant difference found between the samples regarding the second salient feature. Both samples performed similarly in that they both did not recognise that a search engine was a database and not the whole Web. A number of the illustrations provided by users had a computer and an arrow moving from the computer to a giant globe (the Web). Graham and Metaxas (2003) found that people use the internet with little or no understanding of how it actually works. This finding can be extended to search engines as well. While they have become an integral part of internet usage, it was still surprising that the two samples had similar conceptions of this particular aspect of search engines but it was well in line with observations made that very few people really have a coherent concept of cyber-space (Graham & Metaxas, 2003).

There was no significant difference regarding participantsô recognition that search engines use specific algorithms to collect their databases. In both samples there was very little understanding of this. This result is not surprising as it is a more technical aspect of search engines that most users are unlikely to even be aware of as their only interest is having their queries resolved (Spink et al., 2006).

There was no difference between the samples regarding the identification that search engines look for terms/phrases and related term phrases or extensions. Indeed, this is something that not a single participant in both samples was able to identify. This result is expected as most people have no real concept of how cyber-space and the systems therein found truly operate; a fact that spans the era of library catalogue systems (Borgman, 1986). Less transparent search engine functions such as triangulation based on extensions is difficult for even the most competent user (Liben, 2001). This particular function is probably only really understood by system designers as principles behind the definition of an ñextensionò (i.e. retrievable information related to a particular search) most likely vary from system to system. That is, because programmers/designers might interpret an extension differently, this would influence the manner in which they compute this function and the different results that would be produced (consider different órelevanceô principles across systems) (Jansen & Spink, 2006). Participants sometimes commented ñhow is this even relevant; it doesnôt make senseò. Even with Googleôs ñfind similarò feature for instance, these comments could be seen as indicating confusion related to the more abstruse mechanisms behind functional aspects of search engines. This particular salient feature is highly technical in its functioning and it was expected to be beyond the grasp of most participants. This result indicates just how removed some aspects of search engines remain (Graham & Metaxas, 2003).

There was no significant difference between the samples regarding the fact that different search engines have different ranking algorithms. This may seem like a strange result considering the fact that Spink et al. (2006) found clear and seemingly obvious differences in the way different search engines rank results for similar queries. But considering both Googleôs dominance and its tunnel-visioning impact on the 2010 sample; as well as the apparent distrust in even bothering to pay serious attention to the results, let alone their ordering system for the 2000 sample; it is not surprising that for very different reasons, neither of the samples were able to significantly identify this particular fact about search engines. Indeed, a number of participants in the 2010 sample were able to identify that Google has a ranking system but they could not imagine other search engines having the same feature, and further posit that they might be different.

For the last salient feature, a significantly greater awareness was present in the 2000 sample pertaining to the fact that search engines display a hyperlink to the original location of the information sought. This difference might be explained with reference to the lack of interest participantsô in 2010 had about hyperlinks. Their sole concern was in fact the number of matches their search terms had with identifiers on the webpage or database (i.e. extracts (blurbs) and rankings). As information retrieval is much faster now, the need to even bother inspecting the hyperlink was probably completely removed for the 2010 sample. Indeed, only a single participant indicated an understanding of this salient feature in the 2010 sample. In the 2000 sample a significant awareness of this feature could be explained by participantsô need to verify information retrieved. An integral part of that process might have been to do so by tracing the information right to its source.

While seven of the seventeen salient features were more frequent at time 1 (2000) and the same number of features were more frequent at time 2 (2010) there have been clear improvements regarding the completeness and accuracy of search engines.

Firstly, the marked improvements in performance by the 2010 sample would indicate that mental models of search engines have improved reasonably over the last decade as all the performance differences were significant. This would indicate that usersô abilities to describe, explain, and predict system behaviour have shown considerable improvement when compared to what they were a decade ago. These tangible improvements give an indication that usersô interactions with the system have truly become more efficient as a function of greater exposure and frequent use; a finding consistent with that of Muramatsu and Pratt (2001).

Regarding the equal distribution of salient features, there has been a clear, significant improvement in the comprehension of the various salient features that make up a search engine a decade later. This is a point that is easily defensible if the chi-squared results are considered along with the cluster analysis results. In effect, where Thatcher and Greyling (2003) identified four clusters, two of which displayed highly incomplete and inaccurate mental models of search engines, only the first cluster of the 2010 sample shared that quality. The remaining two clusters (52 participants in total compared to the 47 in clusters 3 and 4 of the 2000 sample), particularly the third, showed much more accurate and complete mental models of search engines.

The chi-squared comparisons are in many ways an indication of the state of search engines in their respective eras and the demands they placed on their respective users. A large proportion of users in more recent times have gravitated towards Google and they have enjoyed considerable success with it (SEW, 2009). Thatcher and Greyling (2003) commented on how search engines hid a lot of their salient features and thus made it difficult for users to develop effective mental models. It would seem not much has changed in that regard as the number of salient features the respective samples were able to identify, whilst qualitatively different, remained the same in number. This concern is minor and distracts from a far more important point. It would seem that even though search engines still hide a great number of their features, they have configured their design specifications to such a high level that this need has been completely negated (certainly in terms of task performance on this particular task). That is, it would seem that not only have users become faster and more frugal in their search process, but they are also retrieving accurate information more frequently. The salient features that occurred less in 2010 were either more technical or they are simply inconsequential for search performance.

The poor user query formulations noted by a number of authors (e.g. Slone, 2002; Spink et al., 2000; 2001; 2006) which allowed them to infer that inaccurate information was consistently being retrieved (an inference without concrete evidence) was not demonstrated as a concern because of the more user-friendly, if still fairly disingenuous nature of search engines. In fact, users did not seem too concerned about how search engines work and their more nuanced technical aspects. It would seem that against the recommendations made by Thatcher and Greyling (2003), search engines remain mysterious and hinder effective user formulation of more complete and accurate mental models. However, search engine designers have made modifications (e.g. crawler driven by ‗relevance' principle, drop-down menus, larger databases, automatic search term/phrase correction feedback, etc) that have meant users get the most out of them without comprehending much about their mechanics.

4.8 Summary of the practical and theoretical implications

There was a significant difference in performance between the two samples even though the two samples had an equal number of salient features. There was one less cluster in the 2010 sample and in general, the clusters found in 2010 indicated that an improvement in the mental models of search engines was apparent, even if, on a more technical level, these mental models remained still quite inaccurate and incomplete.

There are a number of implications both practically and theoretically here. Firstly, it would seem that having a better mental model of a particular system (even if only moderately better) does in fact help in improving performance in using that system. This is one of the key conditions found in the definition of mental models (Borgman, 1986; Staggers & Norcio, 1993). This result highlights the importance of developing the mental models of users a great deal more in attempting to assist them to effectively interact with a system. It is a result that holds great value for organisations in particular, as it provides guidance as to the potential benefits of using the development of accurate and complete mental models as a training outcome.

The results also suggest that systems can be designed to negate the need for highly accurate and complete mental models as well. The mental models held by users of the 2010 sample, whilst reasonably better than those of the 2000 sample, still were not close to being as accurate and complete as one would hope for effective interaction with a search engine. Even so, the modifications search engines have made and the apparent user-friendly tools they have employed have clearly allowed for significantly improved performance in executing a directed search task. The marked reduction in steps required to execute and resolve a query were significantly reduced as a result, largely, of the modifications that search engines now display (e.g. drop-down search menus, highlighted extracts, predictive search function, etc). The time taken was also significantly reduced because of these modifications, even though increased internet speeds were also a large contributor in this regard. The most obvious practical implication of this particular result is that a clearer understanding of a userôs basic mental model of a system could aid a system designer in developing a system that is able to identify and appeal to the most important aspects of usersô capabilities and search needs. That is, highly sophisticated system designs can overcome the need for systems to reveal their operating mechanics if they have functional aspects that overcome those needs (e.g. assist menus, automated corrective modes, predictive search error mechanisms, etc).

Considering that systems can seemingly overcome the need for effective mental models to be developed for optimised system use there is a theoretical question concerning the real importance of mental models in effective system use. A large body of literature spanning multiple disciplines has consistently proposed that fairly accurate and complete mental models are essential for effective system engagement (Doyle & Ford, 1999; Senge, 1999; Slone, 2002; Staggers & Norcio, 1993). There have even been suggestions that anything short of this could mean that systems remain effectively inoperable; a dramatic interpretation some authors have of the extent to which inaccurate and incomplete mental models affect system usage (e.g. Borgman, 1986; Doyle & Ford, 1999). Some of the results prior to this study provided an indication that this belief may have some truth to it. Indeed, the performance of the sample of 2000 would also give credence to that particular argument. However, the marked differences in performance, even in the presence of the relatively modest improvements in mental models of search engines, indicate that this is not necessarily true. It would seem that designs can transcend the need for users to know a large number of aspects concerning how they work and simply provide assisting tools in optimising usersô use of a system. So perhaps it is possible that high level mental model formulation is not necessary for effective system use, at least regarding directed search tasks. Thus theoretically, the importance of completely accurate mental models may need to be revised, or at the very least be presented more modestly as there are indications that popular beliefs held about them may not necessarily be valid.

The results, broadly speaking, do show that usersô mental models have aligned a lot better with designersô conceptual models over time. It is clear that this improved alignment might well be credited to designers as they have seemingly overcome the need to develop systems that aid effective mental model formulation for effective system use. In that regard, the results are confusing as they both confirm the theoretical premise of mental models improving with greater system exposure, but also undermine their importance in the presence of highly sophisticated and intuitive systems that seemingly increase performance in the presence of mental models that, improvements considered, still remain largely inaccurate and incomplete.

4.9 Limitations and directions for future research

The current study was intended to match its sample with the study conducted by Thatcher and Greyling (2003) and whilst a close match was achieved, it was not identical as the difficulties in finding the right number of willing participants from the required sample ócategoriesô (i.e. students, scholar, professionals etc) proved extremely difficult. As such, while highly relevant and appropriate, the match between the two samples could have been even closer still. Tracing the same participants used in 2000 would not have been possible as they were anonymous; the likelihood that they would want to participate was anticipated to be low; and some of them may have probably relocated outside of Johannesburg.

As the directed search task was internet based, at times the erratic nature of the connection was such that certain participants suffered from lag and therefore did not experience the same quality internet connection as other participants. Also, given the demands to match the sample as closely as possible, the variable speed of 3G used in some locations (e.g. I.T Company, Law firm and Hospital) meant that the speeds experienced by the participants were not exactly the same on a few occasions. It would be recommended that a location with a good level of stability regarding internet speeds be used in future studies and that an attempt to ensure the whole sample performs the task in this location is made. On a practical level this is a very difficult demand to come to terms with and a great deal of planning would be required.

Another consequence of matching the sample from 2000 was that its limitations were inherited. That is, the sampleôs demographic characteristics could have been a lot more evenly distributed in terms of gender and race amongst others. The sample was dominated by males and the most prominent race was white. Such an endeavour in future studies would allow for differences on a number of demographic characteristics to be assessed and may prove useful in the South African context in particular. This may provide a platform for interesting findings in this regard.

Another limitation of this study is that it was correlational in nature and therefore no causal conclusions could be made. An effort to create conditions that allow for causal conclusions to be made might be a consideration, and given the theoretical assertions concerning mental models, it is not completely beyond reason that such an endeavour would be possible. For instance, the design could be focused on the impact of system exposure on mental models. A control group consisting of people who are not given the manipulation/intervention (system exposure) could be compared with a group (experimental) that is given the manipulation. However, since system exposure is a naturally occurring variable, it will be difficult to find a randomly assigned ócontrolô group of this type. The study could take on a pre-test post-test design. The impacts of the intervention could be measured longitudinally with the post-test being conducted at reasonably later stage. Participants of varying experience with the system could be randomly assigned into one of the two groups and tested (pre-test) to provide relevant baseline measures. This would be followed by the intervention and post-test. All the conditions for testing causality (i.e. manipulation, a control group and random assignment) would be present. Aspects such lateral thinking in system usage could be factored into the performance measure as well. The current studyôs design meant that only modest correlational conclusions were possible.

As most participants in the 2010 sample chose to use Google, it would be worthwhile conducting a study where different search engines are used by a much larger sample to assess whether there is a possibility that different search engines result in different mental models being formed and why this might be the case.

Finally, it would be interesting if a study was conducted on the accuracy and completeness of mental models of the most user friendly systems known in a particular domain. The results of the current study seem to suggest that specific aspects of system design can make the usersô need for more accurate and complete mental models unnecessary in aiding effective system uses so such an endeavour could provide interesting information.

A few theoretical questions arise from this study. How important are mental models? Should the emphasis be on developing system designersô conceptual models of usersô mental models to ensure systems are designed to account for all user weaknesses? Is such an endeavour possible in all design domains? Because the application of mental models is relevant in almost all contexts, these are worthwhile theoretical questions that can provide a meaningful basis for future investigations.

Practical questions stemming from the studyôs findings include what the best way of operationalising the possible advantages of more complete and accurate mental models would be. For instance, how might organisations implement real-world training packages that leverage the knowledge about mental models? Also, how could search engine designs further improve their design interfaces to further facilitate more accurate searching? The practical questions gravitate towards implementing effective mechanisms for getting the most out of mental models in the real world context. As mental models fall squarely in the realm of applied cognitive science, the most salient questions about them are application-based. How can they be leveraged? When do they require modification? What are reasonable expectations regarding the extent to which mental models can be leveraged?

4.10 Conclusion

This study found that mental models of search engines have moderately improved in the last ten years. This improvement was attributed to the significantly better performance by the 2010 sample and the more complex mental model clusters that emerged. It was apparent that the conceptual models of search engines designers hold may have been the underlying reason why a great improvement in performance was found. This observation was confounded by the chi-squared results that indicated that an identical number of salient features of search engines was apparent across the respective samples, suggesting that while improvements in mental models of search engines were found, the nature of the search engine systems and the speed and structure of the internet may also have played a greater role in improved performance than just the moderately more accurate and complete mental models users held.

This studyôs findings provide a great opportunity for the theoretical discourse on mental models to be interrogated further and allows for exciting possibilities in leveraging this knowledge to assist in system design in an ever evolving technological environment.

Reference List

Baeza-Yates, R. & Riberio-Neto, B. (1999), Modern Information Retrieval, ACM Press, New York, NY.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems, 30, 379–388.*

Booth, P., & Brown, G.M. (1989). The Organizational Impact of Computer Systems. In: P. Booth, *An introduction to Human-Computer Interaction.* Lawrence Erlbaum Associates Inc., New Jersey

Borgman, C.L. (1986). The userôs mental model of an information retrieval system: an experiment on a prototype online catalogue. *International Journal of Man-Machine Studies, 24, 47-64.*

Caffarella, M. J., & Etzioni, O. (2005). A search engine for natural language applications. *Proceedings of the 14th international conference on World Wide Web*

Craik, K. (1943). *The Nature of Explanation*. Cambridge, Cambridge University Press.

Doyle, J.K., & Ford, D.N. (1999). Mental models concepts revisited: some clarifications and a reply to Lane. *Systems Dynamics Review, 15, 411-415.*

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

Graham, L., & Metaxas, P.T. (2003). ñOf course itôs true; I saw it on the internet!ò Critical thinking in the internet era. *Communications of the ACM, 46, 70-75.*

Hirtle, S.C., & Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory and Cognition, 13, 208-217.*

Hodkinson, C., Kiel, G. & McColl-Kennedy, J.R. (2000). Consumer web search behaviour: diagrammatic illustration of wayfinding on the web. *International Journal of Human-Computer Studies, 52, 805-830.*

Huck, S. W. (2004). Reading Statistics and Research (4th Ed.). Boston: Pearson

Jansen, B.J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management, 42, 248-263*

Johnson-Laird, P.N. (1989). Mental models. In M.I. Posner (Eds), *Foundations of Cognitive Science.* Cambridge, MA: MIT Press.

Johnson-Laird, P.N. (1983). *Mental Models. Towards a Cognitive Science of Language, Inference, and Consciousness.* Cambridge: Cambridge University Press

Kehoe, C., Pitkow, J.E., & Rogers, J. (1998, July 1998). *GVU's Ninth WWW user Survey*. [Online] Retrieved 3 March 2010 from the WWW: http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-04/.

Kehoe, C., Pitkow, J.E., Sutton, K., Aggarwal, G. & Rogers, J. (1998 October). *Results of GVU's tenth worldwide web user survey*. [Online] Retrieved 3 March 2010 from the WWW: http://www.gvu.gatech.edu/user_surveys/survey_1998_10/.

Kerr, S.T. (1990). Wayfinding in an electronic database: the relative importance of navigational cues vs. mental models. *Information Processing and Management, 26, 511-523.*

Koshman, S., Spink, A., & Jansen, B.J. (2006). Web searching with the Vivisimo search engine. *Journal of the American Society for Information Science and Technology, 57, 1875–1887.*

Liaw, S., & Huang, H. (2006). Information retrieval from the World Wide Web: a user-focused approach based on individual experience with search engines. *Computers in Human Behaviour, 22, 501-517.*

Liben, L.S. (2001). Thinking through maps. In M. Gattis (Eds), *Spatial Schemas and Abstract Thought*. Cambridge, MA: MIT Press.

Montgomery, A., & Faloutsos, C. (2001). Identifying web browsing trends and patterns. *IEEE Computer, 34, 94–95.*

Morgan, G. (1997). *Images Organization (2nd Ed.)* Sage: London.

Mowshowitz, A., & Kawaguchi, A. (2005). Measuring search engine bias. *Information Processing and Management, 41, 193–1205.*

Muramatsu, J., & Pratt, W. (2001). Transparent queries: investigating usersô mental models of search engines. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 217-224, New Orleans, Louisiana.

Norman, D.A. (1988). *The Psychology of Everyday Things*: New York: Basic Books.

Oleson, K E., Sims, V. K., Chin, M. G., Lum, H. C., Sinatra, A. (2010). Developmental human factors: childrens mental models of computers. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 54, 1450-1453*

Rose, D.E. (2006). Reconciling information-seeking behavior with search user interfaces for the web. *Journal of the American Society for Information Science and Technology, 57, 797-799.*

Rouse, W.B., & Morris, M.M. (1986). On looking into the black box: prospects and limits in the search for mental models. *Psychological Bulletin, 100, 349-363.*

Rouse, W.B., & Morris, M.M. (1989). Human information seeking and design of information systems. *Information Processing and Management, 20, 129-138.*

Rutherford, A., & Wilson, J.R. (1992). Searching for mental models in human-machine systems. In Y. Rogers, A. Rutherford & P.A. Bibby (Eds.), *Models in the Mind Theory, Perspective and Application. London:* Academic Press.

Senge, P.M. (1990). *The Fifth Discipline*. New York: Currency Doubleday.

SEW Staff, (2009). *Top Search providers for August 2009*. [Online] Retrieved 20 March 2010 on the WWW: http://searchenginewatch.com/3634991.

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, 1152-1169.*

Speckman, A. (2010, March 17). Web: South Africa has fourth highest Internet usage in Africa. *Business Report*. Retrieved from http://www.busrep.co.za/index.php?fArticleId=5394996&fSectionId=612&fSetId=662.

Spicer, D.P. (1998). Linking mental models and cognitive maps as an aid to organizational learning. *Career Development International, 3, 125-132.*

Spink, A., Jansen, B.J., Blakely, C., & Koshman, S. (2006). A study of results overlap and uniqueness among major web search engines. *Information Processing and Management, 42, 1379-1391.*

Spink, A., Jansen, B.J., & Ozmultu, H.C. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policy, 10, 317-328.*

Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). ñSearching the web: the public and their queries*". Journal of the American Society for Information Science and Technology, 52, 226-34.*

Staggers, N. & Norcio, A.F. (1993). Mental-models: concepts for human-computer interaction. *International Journal of Man-Machine Studies, 38, 587-605.*

Sullivan, D. (2006). *Nielsen NetRatings Search Engine Ratings*. [Online] Retrieved 20 March 2010 on the WWW: http://searchenginewatch.com/2156451.

Thatcher, A. (2008). Web search strategies: the influence of Web experience and task type. *Information Processing and Management, 44(3)*, 1308-1329

Thatcher, A. & Greyling, M. (2003). Mental models of search engines: How do search engines work? In D. Harris., V. Duffy., M. Smith., & C. Stephanidis (Eds) *Human-centred Computing: Cognitive, social and ergonomics aspects.* Mahaw, New Jersey: Lawrence Erlbaum Associates Inc.

Thatcher, A. (2003). Motivation and interests in human information seeking of electronic databases. [PhD thesis]. Johannesburg: University of the Witwatersrand.

Vaughn, M.W. & Resnick, M.L. (2006). Search user interfaces: best practices and future visions. *Journal of the American Society for Information Science and Technology, 57, 777-780.*

Wang, P., Hawk, W.B., & Tenopir, C. (2000), ñUsersô interaction with world wide web resources: an exploratory study using a holistic approachò. *Information Processing and Management, 36, 229-51.*

Ward, J.H. (1963). Hierarchical grouping to optimise an objective function. *Journal of the American Statistical Association, 58, 236-244.*

White, M.D. & Livonen, M. (2001), ñQuestions as a factor in web search strategyò, *Information Processing and Management*, *37, 721-40.*

Whitley, B. E. *(2002). Principles of Research and Behavioral Science. Boston: McGraw-Hill.*

Wilkinson, D.M. & Huberman, B.A. (2007). Cooperation and quality in Wikipedia. *Proceedings of the 2007 international symposium on Wikis,* 157-164, Montreal, Quebec.

Zhang, Y. (2010). Dimensions and elements of people's mental models of an information-rich Web space. *Journal of the American Society for Information Science and Technology,* 61, 2206ï 2218,

Zhang, Y. (2008). Undergraduate students' mental models of the Web as an information retrieval system. *Journal of the American Society for Information Science and Technology,* 59, 2087ï 2098,

APPENDIX

1) Participant information sheet

2) Informed guardian consent

3) Composite WWW experience questionnaire

4) Biographical questionnaire

5) Diagrammatic illustration of a search engine sheet

6) Confidence and satisfaction form

7) Performance mark sheet

8) Retrospective protocol itinerary

Appendix 1: Participant information sheet

Good Day

My name is Sifiso Mlilo and I am currently completing my Masters degree in Industrial/Organisational Psychology at WITS University. I am researching the development of mental models of search engines in the last ten years. You are invited to participate in this study.

Should you choose to participate, you will be required to fill out a short biographical questionnaire along with a composite web experience questionnaire. Following this you will be required to complete a maximum of two computer-based search tasks. While you complete the task(s) your actions will be recorded by onscreen capture software and I will also take note of your actions on a mark sheet. Following the completion of your task, your actions will be played back to you and I will ask you a series of questions related to the actions you undertook in performing your search. After that process you will be required to provide a diagrammatic representation and write a brief description about processes related to the task.

All your personal information, including all the information you provide in completing all the tasks, will be kept completely confidential as no names are asked for at any point during the study. This research study will only discuss general trends therefore no individual's responses will be reported. You will not be advantaged or disadvantaged in any way should you choose to participate or not. The maximum time expected for each participant to complete the tasks is 45 minutes.

Your participation in this research study is completely voluntary. You may choose to withdraw from the study at any time without prejudice. If you choose to participate your actions during the search task will be recorded using a combination of on-screen recording and written notes taken by the researcher. In the event that you are under the age of 18, a parental consent form is provided to ensure parental approval is received for your participation. All the data will be held securely under the care of myself and my research supervisor, Professor Andrew Thatcher, in password protected files for the duration of the research period. Following the completion of the research, all onscreen and textual recordings will be destroyed and only an electronic copy of aggregated data will be kept.

Should you have any queries regarding this research please feel free to contact either myself or my supervisor. The results will be published in a Masters research report, and potentially in a journal article. A summary of the results of the research will be posted on a blog.

Sifiso Mlilo **sifiso.mlilo@gmail.com**

Andrew Thatcher **andrew.thatcher@wits.ac.za**

Appendix 2: Informed guardian consent form

Informed consent form

I ____, the guardian of______________________________ consent for him/her to participate in Sifiso Mlilo's Industrial/Organisational Psychology Masters Research project. I understand that:

- His/her participation will not be seen or heard by any person in this organisation at any time and will be processed by the researcher.
- No direct quotations will be used and complete participant confidentiality and anonymity in reporting will be ensured
- All raw data related to the study will be destroyed after the research is complete.
- No identifying information will be used in the research report.

Signed____________

Appendix 3: Composite WWW experience questionnaire

1. How long have you been using the web: ____________ (years) __________________ (months)

2. What is your preferred browser: ___________________________________

3. How often do you use the WWW: (*please cross one box*)

NEVER	ONCE OR TWICE A YEAR	ONCE OR TWICE A MONTH	ONCE OR TWICE A WEEK	EVERYDAY

4. On average how many hours a week do you spend using the WWW: ______________________ (Hrs)

5. Where do you use the WWW: (*please cross as many boxes as necessary*)

WORK	HOME	FRIEND	PARENT/S	OTHER RELATIVE	INTERNET CAFE

Other: ____

6. How did you learn to use the WWW? ____

What do you use the WWW for? (*please tick as many boxes as necessary*)

UPDATING INFORMATION (I.E. NEWS), SOCIAL NETWORKING	ACCESSING REMOTE EMAIL, NEWSGROUPS, IRC	BANKING, SHOPPING, & OTHER PURCHASES OR TRANSACTIONS
RESEARCH DATA GATHERING	WEB AUTHORING, BLOGGING	TEACHING OR TRAINING
DIRECTED SEARCHING	SOFTWARE/MUSIC DOWNLOADS	GENERAL PURPOSE BROWSING

Other: ____

7. Overall, how would you rate your capability with the WWW (1=NONE through to 5= EXPERT)

1	2	3	4	5

Appendix 4: Biographical sheet

1. What is your job (what are you doing): ____

2. Highest educational qualification: ____

3. Which institution: ____

4. Gender

MALE	FEMALE

5. Home language: ____

6. Racial group:

BLACK	WHITE	INDIAN	COLOURED	OTHER

If other please specify: ____

7. Age: ____

Appendix 5: Diagrammatic illustration sheet

Please provide an illustration of how you think a search engine works:

Please describe briefly in words how you think a search engine works:

____
____
____
____
____

Appendix 6: Satisfaction and Confidence Scale

Please cross the response that best describes how you feel regarding the following statements:

I felt the following regarding **the whole search process**:

COMPLETELY DISATISFIED	SLIGHTLY DISSATISFIED	NEUTRAL	VERY SATISFIED	COMPLETELY SATISFIED

I felt the following regarding **how I got to my answer**:

COMPLETELY DISSATISFIED	SLIGHTLY DISSATISFIED	NEUTRAL	VERY SATISFIED	COMPLETELY SATISFIED

I felt the following regarding **the accuracy of my answer:**

COMPLETELY UNCONFIDENT	SLIGHTLY UNCONFIDENT	NEUTRAL	VERY CONFIDENT	COMPLETELY CONFIDENT

Appendix 7: Performance measure sheet

Participant Number: ___________________

1. Noted number of steps taken to task completion: ____________________________

2. Time taken to task completion: ____

3. Accuracy of the answer: ____

Appendix 8: Retrospective protocol itinerary

Why did you choose the search engine that you did?

Why not use another search engine instead?

Why did you use those search terms? (Ask why search query refined/enlarged/changed where relevant)

Why did you click on that particular link?

Why did you go back?

Why did you open up another browser?

Why did you stop your search here?

Why did you terminate that path?

Note: While these questions are of the sort the researcher will ask all the participants following their completion of the task they are not set in stone. The researcher will ask appropriate questions that relate to the specific actions of each participant where the questions above are not relevant or sufficient.

Appendix 9: Ethics clearance certificate