Evaluating Websites and Web Services:
Interdisciplinary Perspectives on User Satisfaction

Denis Yannacopouloς
*Technological Educational Institute of Piraeus, Greece*

Panagiotis Manolitzas
*Technical University of Crete, Greece*

Nikolaos Matsatsinis
*Technical University of Crete, Greece*

Evangelos Grigoroudis
*Technical University of Crete, Greece*
Chapter 14
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

Hanna Jochmann-Mannak
University of Twente, The Netherlands

Leo Lentz
Utrecht University, The Netherlands

Theo Huibers
University of Twente, The Netherlands

Ted Sanders
Utrecht University, The Netherlands

ABSTRACT
This chapter presents an experiment with 158 children, aged 10 to 12, in which search performance and attitudes towards an informational Website are investigated. The same Website was designed in 3 different types of interface design varying in playfulness of navigation structure and in playfulness of visual design. The type of interface design did not have an effect on children’s search performance, but it did influence children’s feelings of emotional valence and their evaluation of “goodness.” Children felt most positive about the Website with a classical navigation structure and playful aesthetics. They found the playful image map Website least good. More importantly, children’s search performance was much more effective and efficient when using the search engine than when browsing the menu. Furthermore, this chapter explores the challenge of measuring affective responses towards digital interfaces with children by presenting an elaborate evaluation of different methods.

DOI: 10.4018/978-1-4666-5129-6.ch014
INTRODUCTION

There is a trend in digital media for children to design digital products that are ‘cool’ and ‘playful’. Part of taking a ‘playful’ approach in designing digital products for children is creating age-appropriate graphics, or graphics that children can relate to (Meloncon, Haynes, Varelmann & Groh, 2010). In a corpus study of 100 informational Websites for children, we recognized this playful design approach in many of the analyzed interfaces (Jochmann-Mannak, Lentz, Huibers & Sanders, 2012). More specifically, we identified three types of interface design for children, ranging from 1) classical interface design with a classical interaction style and without playful graphics, 2) interface design with playful graphics, but a classical interaction style and 3) playful interface design with playful graphics and a playful interaction style. In this study, we analyzed what the effects are of these different design approaches of an informational Website on children’s interaction with these interfaces and on children’s affective responses towards these interfaces.

The second important objective in this experiment, is to explore the effects of children’s use of a search engine on children’s search performance and affective responses. Conducting an experiment by letting children interact with digital interfaces is a big challenge. However, measuring children’s affective responses towards these interfaces is an even greater challenge, as will be described in this chapter.

THEORETICAL BACKGROUND

Children’s Informational Interface Design

Interactive products for children can be classified in entertainment, educational and enabling products (Markopoulos, Read, MacFarlane & Hoyniemi, 2008). Websites for children as a specific group of interactive products can also be classified in these three genres. Most Websites for children are aimed at entertaining children, for example by providing computer games. For our study with children’s informational Websites, both educational and enabling Websites are relevant, because most informational Websites are educational and search engines that help children in finding relevant information, can be classified as enabling.

Researchers propose some guidelines for children’s Web design (Nielsen & Gilutz, 2002; Meloncon, et al., 2010). Most of these guidelines were tested and validated with children, but many of the guidelines are not specifically aimed at children, and similar to standard Web design practices for adult Websites. In a large corpus study with children’s informational Websites we identified current design conventions for children (Jochmann-Mannak et al., 2012). This study also showed that designers of children’s Websites often follow general Web design guidelines. A closer look at the data in this study did reveal three categories of informational Websites especially designed for children. The first category is a Classic design type in which the layout of the pages is kept minimal and the design is aimed at simplicity, consistency and focus. We called the second category ‘the Classical Play design type’ in which a classic design approach for the navigation structure is combined with a playful, visual design approach. More effort is spent on the design of graphics, colors and games (Meloncon et al., 2010). The third category was called the ‘Image Map design type’ in which no classic Web design characteristics are used. The visual design and navigation structure on the Websites of this type are based on Image maps that incorporate objects or locations that children know from real life or from fiction. Children can explore this tableau of real life or fictional objects, which makes information-seeking a playful experience (Meloncon et al., 2010). This Image map web design can be compared to ‘spatial metaphors’, which can be
employed to visually represent information, using the universe, the solar system, galaxies, and so on through which the user navigates to locate information (Chen, 2006).

In their study to develop a visual taxonomy for children, Large, Beheshti, Tabatabaei, and Nesset (2009) emphasized the importance of movement and color in any visualization designed for children. They argue that “such characteristics do not necessarily influence positively the effectiveness of a taxonomy, but the affective reaction of users, and especially of children, that should never be underestimated. If the presentation is not interesting and fails to catch the attention of users, it is unlikely to invite their repeat visits. It also might be argued that intrinsic to visualization schemes is the ability to provoke interest and even fun” (p. 1818).

Two Search Strategies: Keyword Searching and Browsing

In the beginning of the Internet era a general assumption was made by researchers that browsing-oriented search tools, relying on recognition knowledge, were better suited to the abilities and skills of children than keyword search tools. The argument was that browsing imposes less cognitive load on children than searching, because more knowledge is needed to retrieve terms from memory when searching than simply to recognize offered terms when browsing (Bilal, 2000, 2001, 2002; Borgman, Hirsh, Walter & Gallagher, 1995; Large & Beheshti, 2000; Large, Beheshti, & Moukdad, 1999; Schacter, Chung & Dorr, 1998; Bilal & Watson, 1998).

Schacter et al. (1998) found that with both highly specific and vague search tasks, children sought information by using browsing strategies. In their research on children’s internet searching on complex problems with thirty-two children in the age of 10 to 12 years, they reported: “Children are reactive searchers who do not systematically plan or employ elaborated analytic search strategies” (p. 847).

Bilal (2000) found in her research on the use of the YahooOlligans! Web Search Engine that most of the children (she observed twenty-two children in the age of 12 to 13 years) used keyword search. Only 36% of the searches were performed by browsing under subject categories. This finding may have been affected by the type of search task that was given in this research: a fact-driven query that automatically stimulated children to use keyword search instead of browsing the categories.

Revelle, Druin, Platner, Bederson, Hourcade and Sherman (2002) report on the development of a visual search interface to support children in their efforts to find animals in a hierarchical information structure. To examine searching and browsing behavior, 106 children (aged 5 through 10) participated in an experiment on this visual search interface. The researchers found that: “(…) even young children are capable of efficient and accurate searching. With the support of a visual query interface that includes scaffolding for Boolean concepts, children can use a hierarchical structure to perform searches and construct search queries that surpass their previously demonstrated abilities with the use of traditional search techniques” (p. 56).

By tracking the web logs of The International Children’s Digital Library (ICDL), Druin (2003) found that, of 60,000 unique users between the ICDL’s launch in November 2002 and September 2003, approximately 75% of the searches used category search (browsing), 15% used place search (by selecting a place using a world interface) and just over 10% of the searches used keyword search.

Hutchinson, Bederson and Druin (2006) found that children are capable of using both keyword search and category browsing, but generally they prefer and are more successful with category browsing. They explain this finding in relation...
to children’s ‘natural tendency to explore’. Young children tend not to plan out their searches, but simply react to the results they receive from the Information Retrieval system. Generally, their search strategies are not analytical and do not aim precisely at one goal. Instead, they make associations while browsing. This is a trial-and-error strategy.

It is clear that research results are very diverse when it comes to search strategies used by children. The results seem to depend on the type of interface used in the studies and the type of search task that was given to children. However, the trend in literature is that browsing is more suited for children than using a search engine.

**Difficulties with Keyword Searching and Browsing**

Formulating a search query might be difficult for children, because they have little knowledge to base ‘recall’ on (Borgman et al., 1995; Hutchinson, Druin, Bederson, Reuter, Rose & Weeks, 2005). Besides, for searching relevant documents using keyword search, correct spelling, spacing and punctuation are needed. Children have difficulty with spelling and often make spelling errors (Borgman et al., 1995; Druin, Foss, Hatley, Golub, Leigh Guha, Fails, 2009). That is why an information retrieval system should be able to handle spelling errors, to help children find relevant documents using keyword search. Deciding on a single keyword is also difficult for a child, because children tend to use a full natural language query, especially with complex search tasks (Marchionini, 1989; Druin, et al., 2009). Thus, a system should also be able to handle natural language queries to find relevant information. In a comparison study between children and adults, Bilal and Kirby (2002) found that when children employed keyword search, most of their queries were single or multiple concepts, just like adults do. However, adults employed advanced search syntax, while children did not use this.

Browsing taxonomies may also be difficult for a child, because taxonomies in children’s Web portals such as Kidsclick.org and Dibdabadoo.com use hierarchically structured taxonomies that may impose considerable cognitive load. Only a part of the hierarchy is displayed at any one time, and users must guess which route might eventually take them to the relevant term within the hierarchy (Large, Beheshti, Nesset & Bowler, 2006). With category search (i.e. browsing), children also have trouble finding the right category, because they have little domain-knowledge to decide which category is optimum. In addition, problems with browsing tools are mostly the result of a lack of vocabulary knowledge. Children often have difficulties understanding abstract, top-level headings, because their vocabulary knowledge is not yet sufficient to understand such terms (Hutchinson et al., 2006). Therefore, formulation of headings should be adjusted to children’s vocabulary knowledge, using simple, concrete search terms.

Children may not think hierarchically like adults and may have trouble understanding the way in which hierarchically based categories are constructed. Knowing what their understanding of categories is, can therefore be of great value in designing browsing tools. Bar-Ilan and Belous (2007) tried to understand which browsable, hierarchical subject categories children create by conducting a card sorting experiment with twelve groups of four children in the age of 9 through 11 years. They suggested terms to the children through 61 cards. The children were free to add, delete or change terms. The researchers found that the majority of the category names used by existing directories were acceptable for the children and only a small minority of the terms caused confusion. Finally, often information in browsing systems is alphabetically displayed, requiring good alphabet skills. Many children have problems with alphabetizing and therefore have trouble finding information in such browsing systems (Borgman et al., 1995).
Children’s Search Behavior Characteristics

Bilal (2000) found in her research on the use of the Yahooligans! Web Search Engine that children were chaotic in their search performance: they switched frequently between types of searching (i.e. keyword search or browsing), they often looped their keyword searches and selected hyperlinks, and they frequently backtracked. These findings suggested that children want to combine different search strategies during one search task.

Bilal and Kirby (2002) also found that children were more chaotic in their search performance than adults. In their research, they compared search behavior between twenty-two children (aged 12 through 13) and twelve graduate students. Children made more web moves, they looped searches and hyperlinks more often, they backtracked more often, and they deviated more often from a designated target. The researchers concluded that adults adopted a “linear or systematic” browsing style whereas most children had a “loopy” style. They explain that this “loopy” style can be caused by children’s lower cognitive recall, because the web imposes memory overload that reduces recall during navigation. They also found that children scrolled result pages less often than adults.

We should keep in mind however, that most of these studies were conducted in a time that children did not make use of computers and the Internet as much as they do anno 2013. Children nowadays are much more experienced users of digital interfaces because of iPads, Facebook, online gaming, etc., which makes it difficult to apply these research results to children’s current information-seeking and navigation behavior on digital interfaces.

What we have learned so far from this theoretical background is that playful interface design emerges in the genre of children’s informational Websites. Literature on children’s search behavior and on problems and successes that children experience during information-seeking, especially discusses in pragmatic issues such as query handling and comprehensibility of taxonomies. However, the emergence of playful interface design asks for a broader focus than pragmatic issues. Also hedonic issues of playful interface design should be studied. It is assumed that product characters can be described by two attribute groups: pragmatic and hedonic attributes. Pragmatic attributes are connected to the users’ need to achieve goals (e.g. finding information on an informational Website). Hedonic attributes are primarily related to the users’ self. A product can be perceived by users as hedonic because it provides stimulation by its challenging character or identification by communicating personal values to relevant others (Hassenzahl, 2004). Hedonic issues of interface design will be discussed in the next part of the theoretical background.

Fun and Engagement

From the beginning, research on interaction with digital interfaces is dominated by pragmatic issues such as the utility and usability of these systems (Thüring & Mahlke, 2007). This is the same for research on children’s interaction with digital interfaces (Borgman et al., 1995; Bilal, 2000; Druin, 2003; Hutchinson, 2005). Usability, in particular, is a key concept for capturing the quality of use of digital products in which effectiveness and efficiency of system use is measured. The third component of the usability concept is ‘user satisfaction’. Although this is measured using subjective judgments of users, these are mostly based on efficiency and effectiveness of interface usage.

In the field of Interaction Design for Children (IDC), there is a strong downplay in research about efficiency and task completion (Yarosh, Radu, Hunter & Rosenbaum, 2011). Instead of usability and satisfaction, that are goal related, desirability (being ‘cool’) has become very relevant in the community, which is not goal related. Malone (1980) pioneered the study of fun as an important aspect of software, and published guidelines de-
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

Design for fun (Malone, 1984). According to his constructivist view, children acquire knowledge through experience. But for many years the study of fun in software was of marginal interest. In recent years there has been increasing interest in fun (Read, MacFarlane & Casey, 2002). Yarosh et al. (2011) report that ‘enjoyment’ and ‘fun’ are the most important values in 24% of the papers presented on the yearly Conference on Interaction Design & Children. They even claim that the values enjoyment and fun are that ubiquitous in the community, that they are no longer explicitly discussed, but that they became general assumptions when designing interactive products for children.

User Experience

Also in the general field of Human Computer Interaction (HCI) researchers argue for a broader perspective on user experience (UX) (Hassenzahl & Tractinsky, 2006; Thüring & Mahlke, 2007) which can include, besides perceived usability - beauty, overall quality and hedonic, affective and experiential aspects of the use of technology (van Schaik & Ling, 2008). Considering this broader perspective on UX, the reason for designing playful interfaces for children – as described in the Introduction of this chapter - becomes more clear. Playful design might have a positive effect on children’s overall appraisal of a digital interface. This hypothesis is based on the idea that overall appraisal of a digital product is influenced by perception of both instrumental qualities (for example, effectiveness of a product) and non-instrumental qualities (for example, beauty of a product) as proposed by the Components of User Experience model (Thüring & Mahlke, 2007). Following this idea, playful design might have a positive effect on children’s perception of hedonic quality, because children might feel stimulated by the creative and innovative interaction style or they might be able to easily identify with the playful environment. These are the hedonic attributes of stimulation and identification that are primarily related to the users’ self as described by Hassenzahl (2004). Thüring and Mahlke (2007) propose that both perception of instrumental and non-instrumental qualities have an influence on users’ emotional reactions (such as subjective feelings, motor expressions or physiological reactions), which also has influence on overall appraisal of a system. For example, a slow working system (instrumental quality) may lead to frustration (negative emotion). At the same time, this slow working system might be presented with a creative interface design that may lead to enjoyment (positive emotion). Both experienced emotions have an influence on the overall appraisal of the system.

The Interplay between Components of User Experience

What is interesting to know for designers of digital products is how the overall quality of an interactive product is formed. Evaluating interactive products is very complex, because many factors influence the quality of an interactive product: usability, beauty, overall quality, hedonic quality, and affective and experiential aspects of the use of a product.

Tractinsky, Katz and Ikar (2000), conducted an experiment to test the relationship between user’s perceptions of computerized system’s aesthetic beauty and usability. Perceptions were measured before and after actual use of the system. Both pre and post-use measures indicated strong correlations between perceived aesthetics and usability. Post-use usability ratings were not affected by actual usability (i.e. objective measured usability), which made Tractinsky et al. (2000) conclude that a product’s beauty is a stronger indicator for its perceived usability than its actual usability. In other words, they claimed that “what is beautiful is usable”. Tractinsky et al. (2000) propose the occurrence of a so-called halo-effect. The beauty of an interface overrules all other interface characteristics and therefore influences users’ overall evaluation of the system.
Hassenzahl (2004) also studied the relation between perceived aesthetics and usability. He investigated the interplay between two product evaluations, beauty and goodness and the following perceptions of product attributes: pragmatic quality (i.e. usability as perceived by the user), hedonic quality of stimulation (personal) and hedonic quality of identification (social). He found that beauty as an evaluation was related to the hedonic quality of identification (e.g. a product is perceived as professional, valuable or presentable, etc.). Hassenzahl (2004) found that goodness was more closely related to attributes of pragmatic quality (e.g. a product is perceived as simple, practical, clear, predictable, etc.), especially when participants also interacted with the product under evaluation. These results of Hassenzahl (2004) contradict the results of Tractinsky et al. (2000), because Tractinsky et al. (2000) found no significant main effect of usability on post-use ratings of usability and beauty. In contrast, Hassenzahl (2004) did find an effect of actual usability on perceptions of usability. Hassenzahl (2004) explains this contradiction in results by the fact that Tractinsky’s manipulation of usability was unlikely to induce stress for the participants, which makes any impact on post-use ratings of usability unlikely.

To study how the overall quality or goodness of an interactive product is formed, van Schaik and Ling (2008) also conducted an experiment on the interplay between components of UX. They found that all measures (i.e. evaluation of goodness, attributes of hedonic and pragmatic quality, task performance and mental effort) except evaluation of beauty, were sensitive to manipulation of web design. Evaluation of beauty was influenced by hedonic attributes (identification and stimulation), but evaluation of goodness was influenced by both hedonic and pragmatic attributes as well as task performance and mental effort. Attributes of hedonic quality were more stable with experience (i.e. using the interactive product) than attributes of pragmatic quality. Evaluation of beauty was more stable than evaluation of goodness.

Hartmann, Sutcliffe, and DeAngeli (2008) found a link between aesthetics and usability. When users’ usability experience was poor, positively perceived aesthetics could positively influence overall appraisal of a system, suggesting that “aesthetics could be an important determinant of user satisfaction and system acceptability, overcoming poor usability experience.” (p. 176). Furthermore, they argued that the relative importance of aesthetics is related to the user’s background and task. When the user’s task is goal-oriented, then usability factors will weigh more than aesthetic considerations. When the user’s task is action-oriented (the experience is more important than the goal), users choose designs based on a general impression of aesthetics and engagement.

Tuch, Roth, Hornbaek, Opwis and Bargas-Avila (2012) gave an overview of the current state of research on the aesthetics-usability relation. They made a distinction between correlative studies in which aesthetics and usability were not systematically manipulated as independent experimental factors and experimental studies in which these factors were systematically manipulated. The correlative studies showed some evidence for the relation between usability and aesthetics. However, there was only limited inference on the direction of any causality between aesthetics and usability. In the experimental studies, a pure “what is beautiful is usable” notion was only partially supported. Tuch et al. (2012) reported that there was also some evidence that in certain cases the relation is best described as “what is usable is beautiful” (p. 1598). Tuch et al. (2012) conducted an experiment on the aesthetics-usability relation and also found under certain conditions evidence for the relation “what is usable is beautiful”. They found that the frustration of poor usability lowers ratings on perceived aesthetics.

Note that in none of the discussed studies so far children were involved. However, Hartmann et al. (2008) suggest that a metaphor-based interface
style, such as an Image map Website type (considered more aesthetically pleasing and engaging), would be better for children than a menu-based style if they were interacting with it in their leisure time. They were undecided about this if the interface was to be used in the classroom within a formal educational context. Their reason for this assumption was that a metaphor-based style would likely prove more engaging but perhaps at the expense of usability.

In a study to validate the Fun Toolkit, a tool to evaluate technology with children, Sim, MacFarlane and Read (2005) did try to relate the constructs ‘fun’ and ‘usability’. They report that children experience less fun when there were more usability problems. They conclude that it is not all about fun for children and that usability does matter to them.

We now learned that conducting research on hedonic issues such as fun and engagement became more important in the field of child-computer interaction in recent years. The study of pragmatic and hedonic issues of interface design with adult users from a subjective user-centered perspective on quality of use, is called ‘user experience’ (UX). An important topic in this field of research is the interplay between components of UX, such as usability, beauty and goodness. We think that this topic is also relevant for our research on children’s informational Websites and especially concerning the emergence of playful interface design in this genre. However, methods used in studies with adults are mostly not suitable for studies with children. Current applied methods should be reflected on whether they are suited for children. Therefore, we will now discuss literature on the methods used in this field of research with adults and the methods used to measure hedonic components of UX with children.

Methods to Measure the UX Components

As mentioned before, research on interaction with digital interfaces was dominated by pragmatic issues such as the utility and usability of these systems. Methods to measure usability are measuring effectiveness (the accuracy and completeness with which specified tasks can be conducted in a particular environment), efficiency (for example, the amount of time or digital events required to reach a specified goal) and satisfaction in using the system (based on instrumental qualities of the system). The same methods to measure these factors of usability that are validated in research with adults, can be used well in research with children, as we experienced in a prior explorative study on children’s search behavior (Jochmann-Mannak, Huibers, Lentz and Sanders, 2010).

The evaluation of subjective aesthetic preferences and emotional experiences is more difficult to measure than objective usability scores. According to Laarni (2004), this is one of the reasons why these non-instrumental qualities have played a marginal role in human-computer interaction (HCI) research. Particularly with children, measuring subjective, non-instrumental qualities of a system, for example, with a survey method by asking children to rate product evaluations (such as fun, beauty and goodness) is very difficult, because of risks of satisficing, children’s tendency to say yes irrespective of the question and children’s tendency to indicate the highest score on the scale when scales are used to elicit opinions about software (Markopoulous et al., 2008).

Horton, Read and Sim (2011) report a study in which the reliability of children’s responses on a pictorial questionnaire is tested by asking the same questions about children’s technology twice one week after the other. None of the children
produced the same results for a question after one week, which proves the difficulty of using survey questions with children and the issues with the validity and reliability of questionnaire answers given by children.

Read, MacFarlane and Casey (2002) developed the Fun Toolkit to measure children’s opinions of technology which reduces the mentioned risks of evaluating products with children. The Fun Toolkit consists of four tools: a Funometer, a Smileyometer, a Fun Sorter and an Again-Again table. The Toolkit has been validated in several studies with children (Read & MacFarlane, 2006; Sim, MacFarlane & Read, 2006; Read, 2008).

Visual Analogue Scales (VAS) are often used in survey studies with children. With a VAS, children can identify and visualize their answers, feelings or opinions through pictorial representations instead of textual labels (Markopoulos, et al., 2008). The Smileyometer (Read et al., 2002) is an example of a VAS. But even with such a simple question-answering style as VAS, a child still needs to understand the question, needs to recall relevant information from memory, needs to decide what response is appropriate, needs to translate this response by deciding which pictogram from the VAS is relevant, and requires to physically act to make the selection. All of these steps put high cognitive load on children’s working memory that can be problematic for a child.

An often used subjective method that uses VAS to measure emotional valence and arousal is the Self-Assessment Manikin (SAM) (see Figure 6 in the Method section), developed by Lang (1980). The SAM is a non-verbal pictorial assessment technique that measures the pleasure, arousal, and dominance associated with a person’s affective reaction to a wide variety of stimuli (Bradley & Lang, 1994) which is based on the dimensions of valence, arousal and dominance (Russell, 1980). The use of SAM with children was validated by Greenbaum, Turner, Cook and Melamed (1990).

To measure users’ perceptions of three product attributes (pragmatic quality, hedonic quality–identification and hedonic quality–stimulation) and two product evaluations (beauty and goodness), Hassenzahl, Burmester and Koller (2003) developed the AttracDiff 2 questionnaire. Each of the three product attributes is represented by seven 7-point semantic differential scales (e.g. professiona–amateurish) and the two product evaluations by one 7-point semantic differential scale each (e.g. good–bad). To the best of our knowledge, the AttracDiff 2 questionnaire is not yet validated in survey studies with children.

RESEARCH QUESTIONS

As we learned from the literature, there is a general assumption that digital products or systems for children should be fun and engaging (Yarosh et al., 2011). Non-instrumental product attributes - like beauty and fun- are expected to have an influence on overall appraisal of a digital product, just as instrumental product attributes such as effectiveness and learnability (Thüring & Mahlke, 2007; Van Schaik & Ling, 2008).

It seems that designers of children’s interactive products make their products fun and engaging by adding playful design characteristics, both in visual design and in navigation design (Jochmann-Mannak et al., 2012). Therefore, we are interested whether the adding of playful design has a positive influence on children’s perceptions of hedonic quality and on their overall appraisal of the system. We are also interested in the relation between children’s perception of hedonic quality with their perception of usability and actual task performance.

In a previous explorative study, we did find a positive influence of playful interface design on children’s perceptions of hedonic quality, measured by observing children’s emotional expres-
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

sions while working with informational Websites in a school setting (Jochmann-Mannak et al., 2010). However, most emotional expressions were based on pragmatic product attributes instead of hedonic product attributes. From this study, we concluded that playful design does not have a large influence on children’s overall appraisal of informational interfaces. Usability seems much more important for children than non-pragmatic qualities, such as fun and beauty. However, that study was not set up as an experiment. Therefore, in this follow-up study, we want to test these initial results under controlled experimental conditions.

These are the research questions in our study:

1. What are the effects of different interface design types on…
   a. What are the effects of different design types of informational Websites on children’s search behavior?
   b. What are the effects of different design types on children’s attitude towards informational Websites (emotions, perceptions of pragmatic and hedonic quality and product evaluations, such as beauty, goodness and fun)?
   c. What is the relation between performance (objective usability) and attitude (subjective user-centered perspective on quality of use)?
   d. Is there an interaction between different interface design types with the chosen search strategy (searching with the search engine or browsing the main categories)?

2. If any, what problems and successes do children experience when searching with a search engine or when browsing main categories?
   a. Do these problems and successes relate to design characteristics of the different design types?
   b. Do these problems and successes relate to the quality of the search engine?
   c. Do these problems and successes relate to characteristics of the children?

3. Are existing methods to measure feelings and perceptions of pragmatic and hedonic quality that are used in research with adults also suited for research with children?

RESEARCH METHOD

Experimental Design

The experiment used a 3 x 2 between groups design with two factors: interface design and use of the search engine. Three versions of the same Website varied in aesthetics and navigation style. We did not want to have an influence on children’s natural search behavior by telling them to search by using the search engine or to browse by using the navigation to find information. Therefore, the use of the search engine was manipulated by presenting the Websites with or without a search engine. For each of the three Web design conditions, half of the children used the search engine and half of the children did not use the search engine.

Both independent variables were between-subjects; each child participant used one of the three interface designs and did (at least for one of the search tasks in the experimental session) or did not make use of the search engine at all. Outcome measures included perceptions of product attributes (pragmatic and hedonic quality), evaluations of the Websites (beauty, goodness and fun), objective performance measures, subjective emotion measures and objective emotion measures.

Participants

There were 158 children in the age of 10 to 12 years old that took part in the experiment (70 boys and 88 girls, see Table 1a), with an average
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

age of 10.80 (SD = .65). From these children, 67 were in the fifth grade and 91 were in sixth grade. The children were tested on four different primary schools in the Netherlands. Children were randomly assigned to the three types of Websites concerning age, gender, school, grade and experience with the internet, based on randomization checks. The number of children that used the three different Website versions and the number of these children that did or did not use the search engine are presented in Table 1b.

**Materials and Equipment**

For the manipulation of the interface design, an existing online encyclopedia for children was used (i.e. Junior Winkler Prins online encyclopedia). By using a fully working, existing Website, ecological validity is higher than by building a prototype Website for the experiment. The disadvantage of working with an existing Website is that the Website comes with real life flaws. For example, the search engine on the Website in our experiment did not provide query suggestions or spelling suggestions. Besides the existing version of the Website, two other fully working versions of this same Website were created; each of the three versions with a different interface design. Further, for each of the three versions, again there were two versions: one version with and one version without a search engine. In total, there were six different versions of the same Web site in the experiment.

The Classic version was presented with classical aesthetics and with a classical navigation style (see Figure 1). The Classical Play version (see Figure 2) was presented with expressive aesthetics, but with a classical navigation style. The Image Map version (see Figure 3) was presented with both expressive aesthetics and a playful navigation style. The three Website versions were identical concerning the main categories and subcategories, the menu structure, the content and the logo, to control for effects of these factors. Also, on the deepest navigation level, all three Website versions referred to the same target pages with the same layout for each of the three versions (see Figure 4).

The experiment ran on a laptop (Intel Core, 2.27 GHz, 4.0 GB RAM, Microsoft Windows 7 operating system) with a remote 20” monitor that the children worked on (Figure 5). The screen activities were recorded with Morae usability software (Techsmith) and video and audio recordings were made with a webcam. The children filled out an online questionnaire developed in PX Lab, an open source collection of Java classes and applications for running psychological experiments (Irtel, 2007). Childrens’ electrodermal responses (physiological measure of emotional arousal) were measured with a Q Sensor (Affectiva) (Poh, Swenson, Picard, 2010).

<table>
<thead>
<tr>
<th>Website Version</th>
<th>Number of Children that Used the Search Engine</th>
<th>Number of Children that did not Use the Search Engine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classic version</td>
<td>26</td>
<td>25</td>
<td>51</td>
</tr>
<tr>
<td>Classical play version</td>
<td>26</td>
<td>26</td>
<td>52</td>
</tr>
<tr>
<td>Image map version</td>
<td>25</td>
<td>30</td>
<td>55</td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>81</td>
<td>158</td>
</tr>
</tbody>
</table>

Table 1. (a) Age and gender of the participants. (b) Distribution of children over Website versions and use and non-use of the search engine
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

Data Collection

Measuring Performance

Each child conducted the same five fact-based tasks (see Appendix 1 for the full task descriptions). We tested ten tasks in a pilot test with a group of 14 children. Based on the results of this pilot test, we selected five tasks for the final experiment that varied in difficulty of conducting the task with the search engine or by browsing the categories. The task about Columbus, for example, was difficult to conduct both by using the search engine or by browsing the categories, because children had to find out which nation had discovered America 500 years before Columbus discovered America. The answer to this question was not mentioned on the content page about
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

Figure 3. Image map Website version

Figure 4. Example of a target page that is identical for all three Website versions

Figure 5. Experimental setting
‘Christopher Columbus’, but only on the content page about ‘Discoveries’ or at the content page about the ‘Vikings’.

Task performance was measured by logging the amount of time and clicks needed to conduct the tasks. Also, per task was analyzed whether the relevant Web page was found and if and - if so - what type of help was offered. Recordings of the screen activities and video (see, for example, Figure 10) and audio recordings of the children were qualitatively analyzed, for example, to indicate what problems children experienced with particular design characteristics.

Measuring Affectivity

The children gave responses to an online questionnaire to measure their feelings, that consisted of three parts: 1) the Self-Assessment Manikin (SAM) (Lang, 1985) to measure children’s valence and arousal concerning the Website versions, 2) an adaptation of the AttracDiff 2 (Hassenzahl et al., 2003) to measure children’s perceptions of pragmatic and hedonic quality and 3) a questionnaire to measure children’s evaluations of beauty, goodness and fun by giving a report mark from 1 (= lowest score) to 10 (= highest score). Each of the items in the questionnaire was presented underneath a picture of the homepage of the Website version that the child had used to conduct the search tasks and that had to be evaluated by the child (for example, Figure 7). An added bonus to this study is that we can also evaluate whether this method is suited to measure product affectivity with children.

In the first part of our questionnaire, we used the same 5-point bipolar scale version of the SAM as used by Greenbaum et al. (1990) (see Figure 6) instead of the original 9-point bipolar scale (Lang, 1980). We decided to only measure the dimensions of valence and arousal and to leave out the dimension of dominance, as Thüring and Mahlke (2007) also did.

For the second part of our questionnaire, we developed Visual Analogue Scales (VAS) based on the AttracDiff 2 questionnaire (Hassenzahl et al., 2003). We translated 15 of the 21 bipolar verbal anchors for the product attribute groups from the AttracDiff 2 questionnaire in bipolar picture anchors as presented in Appendix B (for example, Figure 7: bipolar picture anchor of the semantic differential scale ‘Clear – Confusing’). We used these 15 items in the pilot test and asked the children to explain the meaning of the pictures in the visual versions of the semantic differential scales. We decided to remove four of the items from the questionnaire (marked with an asterisk in Appendix 2), because none of the children could give a meaning to the pictures that was close to the original meaning of the semantic differential scales. We decided to use 5-point scales instead of 7-point scales. Younger children tend to respond in an extreme manner when asked to use Likert rating scales, whereas older children are more capable of providing graded ratings in the middle of the scale. As tasks become more subjective and emotion focused, as is the case in our study, children’s extreme scores, regardless of age, increase (Chambers & Johnston, 2002). Therefore, although the children in our study were between 10 and 12 years old, providing them with more than three ratings in the middle of the scale, would not add value to the rating scales. For younger children, 3-point scales would probably be most suited, because of their tendency to give extreme ratings. However, children in the age of our study (i.e. 10-12 years old) are capable of differentiating between more and less extreme ratings on a 5-point scale. We did not use bipolar picture anchors for the product evaluations of beauty, goodness and fun. Instead of that, in the final part of the questionnaire we asked the children to give report marks for beauty, goodness and fun of the Websites.

Product experience is a multi-faceted phenomenon that involves feelings, behavioral reactions, expressive reactions, and physiological reactions.
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

Figure 6. SAM 5-point bipolar scales for valence (A) and arousal (B)

Figure 7. Example of the bipolar picture version of the semantic differential scale ‘Clear – Confusing’ presented below the homepage of the Website version under evaluation. (The scores entered by placing the yellow cursor somewhere on the scale, were automatically recorded by the PX Lab software. The left end of the scale was scored as 1 and the right end of the scale as 5.)

(Desmet & Hekkert, 2007). We measured feelings using the online questionnaire as described in the beginning of this section. We also made audio and video recordings of the children, through which behavioral and expressive reactions could be observed and analyzed.

However, because of risks of effects of satisficing (children’s tendency to give superficial responses that generally appear reasonable or acceptable), suggestibility (the influence of the interviewer or evaluator on the children’s question-and-answer process) and children’s tendency to give extreme scale ratings (Markopoulos et al., 2008), we also used a more objective method to measure product experience by measuring physiological reactions. We measured children’s physiological emotional arousal with the Q Sensor.

The Q Sensor is a wearable, wireless biosensor that measures emotional arousal via skin conductance (SC), a form of electrodermal activity (EDA) that grows higher during states such as excitement, attention or anxiety and lower during states such as boredom or relaxation. The sensor also measures temperature and activity...
Typically EDA is recorded as skin conductance by applying a direct current (with two silver electrodes) to the skin (i.e. exosomatic method). Central to this measure is the electrodermal response (EDR). The EDR constitutes a sharp rise in the SC value, followed by a slower drop in conductance. For example, a sudden loud burst of noise will result in an EDR 1-2 seconds later, and this is easily visible in the raw data signal. In general, changes in SC are closely linked to activity of the sympathetic part of the autonomic nervous system. Therefore, researchers and practitioners have taken EDA measurements as further operationalization for constructs such as attention, stress, anxiety, workload, pain, and arousal (Noordzij, Scholten & Laroy-Noordzij, 2012). When children emotionally react differently on the three types of Websites, this might be signaled by differences in the number of EDR and the total amplitude of these EDR per minute during task performance between the three Websites. In Figure 8 an example is presented of the output of the Q Sensor, in which can be seen that the device measures three physical properties at the same time: Electrodermal activity, Electrode Temperature and Acceleration.

Procedure

The study was carried out in the fifth and sixth grades of four different primary schools in the Netherlands in the period of September – November 2011. Only children that could hand in a signed consent form by their parents, could co-operate in the study. All children that co-operated in our study filled out a profile survey in the class room in which we asked about their media use, such as their favorite video game, Website or television show, and the amount of time they spent on the Internet, on video games, or on watching television. Half an hour before a child contributed to our study, the test instructor (i.e. the first author of this chapter) put the Q Sensor on the child’s wrist. In that way, the Q Sensor would be accustomed to the child while the child stayed in his class room. After half an hour, the test instructor came back in the class room to take the child to the room in which the experiment was conducted. Before the child started the actual task performance, first, the child was asked to run up and down the stairs three times to activate the Q Sensor. After that, half of the group of children was asked to watch a short animation film (Disney Pixar– “For the Birds”) to record a base line for the Q sensor that

Figure 8. Example of the Q-sensor’s output for Electrodermal Activity, Electrode Temperature and Acceleration
How Interface Design and Search Strategy Influence Children's Search Performance and Evaluation

was the same for all 158 participants. The other half of the group of participants watched the film at the end of the session to control for a potential order effect of watching the film before or after task performance on the experimental Website.

Before the actual task performance started, the test instructor explained that the child would be asked to conduct five search tasks on a Website, because the designer of the Website wanted to know how children feel about the Website and if children can easily find information. The test instructor emphasized that “the child was not tested in this study, but that the Website was tested”. The five search tasks were provided to the child in random order on separate sheets. When the child received the first search task, the test instructor started recording the screen activity and the video and audio recording of the child. The test instructor sat next to the child during the task performance and offered help when appropriate. Avoiding help and social interaction can make the child feel uncomfortable (Markopoulos et al., 2008). We listed the type of help that was offered for each conducted search task. Types of help were help in choosing the right category or sub-category, help in operating the navigation tools, help with spelling or formulation of a search query, or help in finding the right information on a particular content page. The effects of interface design on children’s search performance were corrected for the help that was offered to the children. After the search tasks were completed, the test instructor started the online questionnaire and asked the child to read and answer the questions presented in the questionnaire.

Data Analyses

To measure the effects of interface design and search engine use on children’s Website performance and attitude towards the Website, a multilevel model was constructed. Independent variables were the type of interface design (Classic, Classical play and Image map) and the use (or non-use) of the search engine. Dependent variables were the percentage of children that were successful in finding the right answers to the search questions (i.e. the chance for success in our multilevel model), the time needed, and the number of clicks needed to find the right information. The model also estimates task variance (because one task can be more difficult than another task), between-children variance (because one child can be more or less skilled in searching information) and residual variance (for example, one child can have more difficulties with task 1, whereas another child can have more difficulties with task 2). By measuring these variances, we can estimate the extent to which we can generalize over tasks and children. When we do not take into account this task variance, between-children variance and residual variance, then the probability of falsely rejecting the null hypothesis is greater than 0.05 (Snijders & Bosker, 2012).

RESULTS

Reliability and Validity of the Affective Response Questionnaire

Before we report the results concerning the effects of interface design and search strategies on children’s search performance and affectivity towards the interfaces, we will first report about the reliability and validity of the used methods.

We measured actual perceptions of hedonic and pragmatic quality using an instrument that was derived from the AttracDiff 2 questionnaire as composed by Hassenzahl et al. (2003). Cronbach’s alphas for the two clusters of bipolar verbal anchors for the constructs of hedonic and pragmatic quality are shown in Table 2. A Cronbach’s alpha of 0.6 is usually regarded as the lower bound of an acceptable reliability for experimental purposes. Both clusters measure the underlying constructs in a reliable way (see Table 2). However, the average scores and standard deviations show
that most children chose the center of the scales instead of the extreme scale ratings, which makes the reliability scores less meaningful. Apparently, children chose the safe, neutral ratings (the center) of the scales, for reasons that we will discuss in the following of this section.

The validity of the questionnaire was studied by a qualitative analysis of the recordings of all respondents that filled out the questionnaire. In total, 2054 items were filled out by the 158 children and 151 of these children uttered a verbal or non-verbal interpretation of at least one of the 13 items from the questionnaire. These 151 children uttered 693 verbal or non-verbal interpretations of the items, which is 33.7% of the total items that were filled out in the experiment. The items from the questionnaire to measure beauty, goodness and fun were left out of this qualitative analysis, because none of the children indicated having problems with the meaning of these items. Of the 693 utterances in total, 224 utterances represented correct interpretations of the items (as intended by Hassenzahl et al., 2003) and 330 utterances represented incorrect interpretations of the items. For example, about the emotional arousal SAM-scale, many children gave the following incorrect interpretation: “I think this is about how easy or difficult the Website is”. Another 84 utterances were verbal indications of incomprehension of the meaning of the items (e.g. “I don’t get it.”) and 33 utterances were non-verbal indications of incomprehension (e.g. by frowning). While filling out the questionnaire, the test instructor helped the children when they asked for help, as discussed in the Method section. Most help was offered by asking a counter-question, for example: “What do you think the picture stands for?” Although, only a third of the items represent children’s interpretations of these items, we think that these interpretations can be related to the entire set of items that were filled out in the experiment. The interpretations of the items will be discussed in the next section.

**Table 2. Cronbach’s alphas of the constructs of hedonic and pragmatic quality (using 5-point Likert scales from 1 to 5)**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Scale</th>
<th>N semantic differentials</th>
<th>Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedonic quality</td>
<td>5-point</td>
<td>5</td>
<td>.63</td>
</tr>
<tr>
<td>Pragmatic quality</td>
<td>5-point</td>
<td>6</td>
<td>.72</td>
</tr>
<tr>
<td>All semantic differentials</td>
<td>5-point</td>
<td>11</td>
<td>.80</td>
</tr>
</tbody>
</table>

**Interpretations of the SAM-Scales for Emotional Valence and Arousal**

Most children (138 of the 158 children) did not utter verbal or non-verbal interpretations of the valence scale (see Figure 6A). Children also did not ask many questions concerning this item. Most interpretations of this item were correct and were related to positive and negative emotional feelings, such as “The Website is kind of fun”. Apparently, children understood the meaning of this valence scale with a more or less smiling manikin, which is related to the Smileyometer of the Fun Toolkit (Read et al., 2002). Therefore, we decided that the valence scale was a valid method to measure children’s emotional valence towards a Website in our study.

The SAM scale for emotional arousal (see Figure 6B) caused many more problems than the scale for emotional valence. In total, 85 children uttered an incorrect interpretation of an indication of incomprehension concerning this scale and 32 children uttered a correct interpretation. Most conspicuous interpretations of the SAM-scale of arousal were: “I don’t get it! Why is his belly exploding?” (while he points at the picture for highest arousal), or “I haven’t got a clue. It looks like a fried egg on his belly, or something like...”
that”. Obviously, for most children the meaning of the SAM-scale for arousal was not clear at all and therefore, in our experiment, the arousal scale has proved to be an invalid method to measure children’s emotional arousal towards a Website.

Interpretations of the Bipolar Picture Anchors for Pragmatic Quality

The picture anchors for the semantic differential ‘technical– human’ were most problematic of all pragmatic semantic differential items for the children as can be seen in Table 3. Only one child gave the correct interpretation of the picture anchors. Many children associated the hearts and flowers with hedonic concepts of ‘love’ and ‘fun’, while it was intended as a pragmatic concept. Also, making a direct translation of these pictures to working with a Website was very difficult for the children. The reliability of the pragmatic scale items increased to .72 if this problematic item ‘technical– human’ was deleted from the list.

Almost half of the interpretations uttered on the picture anchors for the concept ‘complicated–simple’ were incorrect. Many children (27) asked for help interpreting these pictures, which also stresses the difficulty of these pictures. Some children gave a literal meaning to the pictures, such as ‘neat or scratchy lines’ and associated this with hedonic concepts. They could not make a translation to the pragmatic concepts of ‘simple’ and ‘complicated’.

The same problem of literal translation of the pictures was the case with the picture anchors for ‘impractical– practical’. Children asked what “tripping over a stone has to do with searching on a Website?” The interpretation ‘easy– difficult’ was often given to these picture anchors.

Although many children seemed to understand the meaning of the picture anchors for clear– confusing, they interpreted the meaning quite literally, by saying that the main and submenus offered many options to choose from. However, they did not give their opinion about whether these options were clear or confusing.

Most children gave a correct interpretation of the picture anchors for ‘cumbersome– direct’. However, often help was asked from the test instructor and 25 children received help by giving them a counter-question or by explaining the meaning of the picture anchors. Many children gave the correct interpretation “whether you can find it directly or with a detour.” However, most of them based their answer on their own performance instead of on the directness or cumbersomeness of the Website.

Table 3. Frequency table of uttered interpretations of the pragmatic semantic differential items (N = absolute number of children that gave the interpretations)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect interpretations</td>
<td>38</td>
<td>28</td>
<td>15</td>
<td>17</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Correct interpretations</td>
<td>1</td>
<td>27</td>
<td>20</td>
<td>24</td>
<td>35</td>
<td>34</td>
</tr>
<tr>
<td>Verbal indication of incomprehension</td>
<td>18</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Non-verbal indication of incomprehension</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>No utterances</td>
<td>90</td>
<td>98</td>
<td>118</td>
<td>112</td>
<td>106</td>
<td>118</td>
</tr>
<tr>
<td>Missing values</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
The pragmatic item that was least problematic for the children was the semantic differential ‘unruly – manageable’. Most interpretations given for this item were correct and little help was offered by the test instructor. The reason for this is that a literal translation of the picture anchors can directly be related to a Website, because a Website can look neat or unruly.

Interpretations of the bipolar picture anchors for perceived hedonic quality

The hedonic items were even more problematic than the pragmatic items (see Table 4), because the interpretations of these concepts were often even more difficult to translate to the use of a Website. The picture anchors ‘easy–challenging’ were given the most incorrect interpretations by the children, that is ‘easy–difficult’. To strictly test the validity of the questionnaire, we decided to score these interpretations as incorrect, because ‘easy–difficult’ is a pragmatic concept as opposed to the hedonic concept ‘easy–challenging’. The reliability of the hedonic scale items increased to .73 if this problematic item ‘easy–challenging’ was deleted from the list.

The picture anchors for ‘cheap–valuable’ caused a lot of problems for the children, because first of all, they did not understand the meaning of the paper hat versus the crown. When the test instructor asked: “What do you think is the difference between the two hats?” children gave the interpretation of ‘poor versus rich’. However, they did not understand how a Website could be ‘poor or rich’.

The children did not understand the picture anchors for ‘amateurish–professional’ at all. They often thought it was about ‘a drilling machine versus a hammer’ and they could not relate these concepts to a Website.

The picture anchors for ‘presentable–unpresentable’ evoked a lot of questions for the test instructor. When children gave the interpretation of an ‘old or new present’, or maybe even a ‘beautiful or ugly present’, they could not relate this ‘present’ to the Website.

Finally, also the picture anchors for ‘lame–exciting’ were problematic for the children. Again, the children had trouble understanding the meaning of the two types of ‘cycling’ to ‘lame and exciting.’ When the test instructor gave help with this first step, most of the time the children gave their opinion about whether searching for information is lame or exciting and not whether the Website was lame or exciting.

Table 4. Frequency table of uttered interpretations of the hedonic semantic differential items

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect interpretations</td>
<td>51</td>
<td>34</td>
<td>33</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>Correct interpretations</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Verbal indication of incomprehension</td>
<td>7</td>
<td>16</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Non-verbal indication of incomprehension</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>No utterances</td>
<td>95</td>
<td>97</td>
<td>113</td>
<td>113</td>
<td>127</td>
</tr>
<tr>
<td>Missing values</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
Distinction between Perceptions of Pragmatic and Hedonic Quality

We also measured the reliability of all semantic differentials together and it turned out that Cronbach’s alpha for all semantic differentials is .80 (see Table 2). It seems that the children perceived all items as the same construct, for example as the construct ‘good or bad’. One child said: “It’s all a bit the same to me” and often children asked for confirmation: “So this means good and this means bad, right?” In other words, the children did not make a distinction between hedonic and pragmatic constructs or between fun and usability.

And although children did not indicate having problems with the report marks for beauty, goodness and fun, it can be expected that these concepts were evaluated as the same construct. This was also reported by Read et al. (2002) concerning the Fun-Sorter in which children needed to sort products by concepts as ‘worked the best’, ‘liked the most’, ‘most fun’ and ‘easiest to use’. They say: “This was quite difficult for the children with the result that some constructs turned out to be quite similar.”

Conclusion Validity Questionnaire Items

From the qualitative analysis of children’s interpretations of the questionnaire items, we can conclude that, although the construct reliability of the questionnaire items is high, the questionnaire is not a valid method to measure children’s perceptions of pragmatic and hedonic quality. The content validity cannot be guaranteed, because often children gave another explanation to the items than was intended by the designers of the questionnaire. Also, construct validity cannot be guaranteed, because items that should measure pragmatic quality, were associated with hedonic quality and vice versa. Besides that, it seems that children do not make a distinction between pragmatic and hedonic constructs at all. Also the SAM-scale for arousal has not proven to be a valid method to measure emotional arousal with children.

The most important problem that children experience with the questionnaire is that they have to interpret the picture anchors and to relate their meaning to using a Website. As children tend to take the picture anchors very literally, translating them to a more abstract concept is very difficult for them. The cognitive load of this task on children’s working memory is too heavy for children to cope with. Furthermore, children tend to relate the picture anchors to their own performance or preferences instead of to the Website under evaluation. Finally, satisficing is a relevant problem in our experiment. Children are prone to satisficing as they find survey participation difficult (Markopoulos et al., 2008), as was the case in our study. From the recordings, we saw that children tend to ask questions about the first four or five items in the questionnaire, but after that fill out the questionnaire very quickly. It is clear that the children gave more or less superficial responses that generally appear acceptable, but without going through all the steps involved in the question-answer process.

Based on the validity analyses of the affective survey question, unfortunately, we can only work with a few items for further qualitative analysis of the affective data. We will work with the results from the SAM-scale for emotional valence, with the semantic differential scale for unruly-manageable and with the report marks for the product evaluations beauty, goodness and fun.

Effects of Interface Design and Used Search Strategy

We will first report the pragmatic effects on children’s search performance of differences in interface design (i.e. Classic, Classical Play and
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

Image map) and the used search strategy (i.e. keyword searching or browsing). After that, we will report the effects on children’s emotional feelings, perceptions of hedonic quality and product evaluations of interface design and the used search strategy.

What are the Effects on Task Performance of Playful Interface Design and Use of the Search Engine?

To establish whether there is a difference between task performance on the three different versions of the Website and between use and non-use of the search engine, the mean percentages for success were compared (see Table 5). Non-use of the search engine occurred in two situations: 1) when the search engine was not offered, children logically could not use the search engine and 2) when the search engine was offered, some children did not use it to search for information. We also compared the mean percentages of finding the right answer when help was offered and when no help was offered by the test instructor and we found significant effects of provided help (see * in Table 5). The following data analyses for the mean percentages for success are therefore corrected for help (mean percentages for help are grey colored in Tables 5, 6 & 7).

No main effect on task performance was observed for the design type of the Websites: there is no significant difference for the percentage of success in finding the right information between the three Website versions ($\chi^2 = 1.02; df = 2; p = 0.31$). However, a main effect was found for the use of the search engine ($\chi^2 = 43.19; df = 2; p < .001$): the percentage of success was much larger when the search engine was used than when the search engine was not offered ($\chi^2 = 27.33; df = 1; p < .001$) and when the search engine was not used ($\chi^2 = 40.63; df = 1; p < .001$). There is no significant difference between the percentage of success when the search engine was not offered and when the search engine was not used ($\chi^2 = 0.41; df = 1; p = .52$). No interaction-effect was found for the use of the search engine and the three Website versions ($\chi^2 \leq 3.04; df = 2; p \geq 0.080$). In other words, the differences between use and non-use of the search engine for success are the same for the three Website versions.

What are the Effects of Playful Interface Design and Use of the Search Engine on Time and Clicks Needed to Conduct the Tasks?

We also compared the mean amount of time and clicks children needed to conduct the tasks between the three Website versions and between the use

<table>
<thead>
<tr>
<th>Website Version</th>
<th>Search Engine Used</th>
<th>Search Engine not Provided</th>
<th>Search Engine Provided, but not Used</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Help</td>
<td>Help</td>
<td>No Help</td>
</tr>
<tr>
<td>Classic</td>
<td>.82 (.53)</td>
<td>.83 (.61)</td>
<td>.63 (.52)</td>
</tr>
<tr>
<td>Classical play</td>
<td>.84 (.62)</td>
<td>.63 (.51)*</td>
<td>.63 (.54)</td>
</tr>
<tr>
<td>Image map</td>
<td>.92 (.42)</td>
<td>.76 (.18)*</td>
<td>.52 (.10)</td>
</tr>
</tbody>
</table>

Note. In all cases, a higher mean score represents a higher percentage for success in finding the right information for the search task. The answers for the binomial success-score (1 = successful, 0 = unsuccessful) are given in Logits that are used for the data analysis (between brackets).

* There is a significant effect of provided help on the mean percentage of success. The percentage of success in finding the right information was significantly lower for the children that used the search engine and received help from the test instructor for both the Classical play Website and the Image map Website ($t \geq 2.01; p \leq .04$).
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

Table 6. Mean time needed in seconds (ln between brackets) using the different versions of the Website, for use and non-use of the search engine and for help provided yes (grey colored) or no

<table>
<thead>
<tr>
<th>Website Version</th>
<th>Search Engine Used</th>
<th>Search Engine not Provided</th>
<th>Search Engine Provided, but not Used</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Help</td>
<td>Help</td>
<td>No Help</td>
</tr>
<tr>
<td>Classic</td>
<td>139.9 (4.94)</td>
<td>247.4 (5.51)*</td>
<td>168.3 (5.13)</td>
</tr>
<tr>
<td>Classical play</td>
<td>149.7 (5.01)</td>
<td>257.0 (5.55)*</td>
<td>141.6 (4.95)</td>
</tr>
<tr>
<td>Image map</td>
<td>112.8 (4.73)</td>
<td>234.6 (5.46)*</td>
<td>170.4 (5.14)</td>
</tr>
</tbody>
</table>

Note. The distribution of the raw data for time was not comparable to the normal distribution. Therefore, we took the natural log of the search times that did show a normal distribution.

* There is a significant effect of provided help on the mean amount of time needed. The amount of time needed to conduct the tasks was significantly higher for children that received help from the test instructor (t ≥ 3.73; p ≤ .001).

Table 7. Mean number of clicks (ln between brackets) using the different versions of the Website, for use and non-use of the search engine and for help provided yes (grey colored) or no

<table>
<thead>
<tr>
<th>Website Version</th>
<th>Search Engine Used</th>
<th>Search Engine not Provided</th>
<th>Search Engine Provided, but not Used</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Help</td>
<td>Help</td>
<td>No Help</td>
</tr>
<tr>
<td>Classic</td>
<td>4.2 (1.44)</td>
<td>7.4 (2.00)*</td>
<td>9.8 (2.29)</td>
</tr>
<tr>
<td>Classical play</td>
<td>4.7 (1.55)</td>
<td>8.1 (2.09)*</td>
<td>7.7 (2.04)</td>
</tr>
<tr>
<td>Image map</td>
<td>3.3 (1.20)</td>
<td>7.2 (1.97)*</td>
<td>9.0 (2.20)</td>
</tr>
</tbody>
</table>

Note. The distribution of the raw data for number of clicks was not comparable to the normal distribution. Therefore, we took the natural log of the number of clicks that did show a normal distribution.

* There is a significant effect of provided help on the mean number of clicks. The number of clicks needed to conduct the tasks was significantly higher for children that received help from the test instructor (t ≥ 2.85; p ≤ .004).

and non-use of the search engine (see Table 6 and 7). Because we also found significant effects of the provided help for these factors (see * in Table 6 and 7), we corrected the data for provided help.

No main effect was observed for the design of the Website versions: there is no significant difference in time needed to conduct the tasks between the three Website versions ($\chi^2 = 2.00; df = 2; p = 0.37$). However, a main effect was found for the use of the search engine ($\chi^2 = 8.10; df = 2; p = 0.017$): less time was needed when the search engine was used than when the search engine was not offered ($\chi^2 = 6.88; df = 1; p = 0.009$) and when the search engine was not used ($\chi^2 = 4.27; df = 1; p < 0.039$). There is no significant difference between the time needed when the search engine was not offered and when the search engine was not used ($\chi^2 = 0.88; df = 1; p = 0.35$).

No interaction effect was found for the use of the search engine and the three Website versions ($\chi^2 ≤ 2.97; df = 1; p ≥ 0.16$). In other words, the differences between use and non-use of the search engine for the time needed to conduct the search tasks are the same for the three Website versions.

No main effect on clicks was observed for the design of the Website versions: there is no significant difference in the number of clicks between the three Website versions ($\chi^2 = 4.08; df = 2; p = 0.13$). However, a main effect was found for the use of the search engine ($\chi^2 = 257.56; df = 2; p < .001$): the number of clicks needed when the search engine was used was lower than when the search engine was not offered ($\chi^2 = 80.27; df = 1; p < .001$) and when the search engine was not used ($\chi^2 = 66.52; df = 1; p < .001$). The number of clicks is also significantly higher when
the search engine was not offered than when the search engine was not used ($\chi^2 = 212.78; df = 1; p < .001$). A reason for this could be that children in the condition without a search engine were normally used to working with a search engine and therefore, were less experienced and needed more clicks to find the information using the navigation structure.

No interaction effect was found for the use of the search engine and the three Website versions ($\chi^2 \leq 3.77; df = 1; p = 0.052$). In other words, the differences between use and non-use of the search engine for the clicks needed to conduct the search tasks are the same for the three Website versions.

In conclusion, children who used the search engine instead of browsing the categories, were more successful in finding the right information and they needed less time and clicks. There were no significant differences for task performance (i.e. success, time and clicks) between the three Website versions. Apparently, interface design of the search environments is less determinant for task performance than the search strategy (i.e. searching or browsing). The differences that we found between the conditions (independently of the used search method; searching or browsing) cannot be assigned to our manipulations of the interface design. These differences should be assigned to the differences between the children that participated in our experiment. In other words, differences in child characteristics, such as their information skills, domain knowledge, operational skills, etcetera, cause more variance in children’s search performance than variation in interface design.

What is the Effect of Interface Design on Emotional Valence and on the Evaluation of Beauty, Goodness and Fun

To test whether there is an effect of design on children’s affective responses, we computed both between groups and within groups analyses of variance. In that way, we took into account the “effect variation” and the “individual variation”. Individual variation is the variation within condition differences called “error”, because we cannot explain the fact that children who were in the same two conditions - who were all treated the same two ways – have different scores. In this way, we also took into account the “subject variation”, which is the variation due to subject variability. For these tests on children’s affective responses, we could only use a limited set of variables that proved to be valid (see Table 8).

There were significant differences for children’s emotional valence and their evaluation of goodness between the three Website versions. The children judged their feeling with the Classical play Website as more positive than with the other two Websites ($F_{2,155} = 3.28; p = .040$). Still, differences for children’s emotional valence and their evaluation of goodness

<table>
<thead>
<tr>
<th>Website Version</th>
<th>Valence</th>
<th>Unruly-Manageable</th>
<th>Goodness</th>
<th>Fun</th>
<th>Beauty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classic (N = 51)</td>
<td>4.12 (.77)</td>
<td>4.25 (.95)</td>
<td>8.66 (1.31)</td>
<td>8.29 (1.56)</td>
<td>8.37 (1.42)</td>
</tr>
<tr>
<td>Classical play  (N = 52)</td>
<td>4.38 (.66)*</td>
<td>4.10 (.79)</td>
<td>8.71 (.84)</td>
<td>8.50 (1.19)</td>
<td>8.86 (1.05)</td>
</tr>
<tr>
<td>Image map (N = 55)</td>
<td>4.05 (.68)</td>
<td>3.90 (.91)</td>
<td>8.22 (1.04)*</td>
<td>8.13 (1.10)</td>
<td>8.42 (1.03)</td>
</tr>
</tbody>
</table>

* There is a significant effect of the type of design on the affective responses.
all scores are between 4 and 5 on a scale from 1 to 5, so the children are very positive about their feelings with all three types of Websites.

The children evaluated the Image map Website as least good ($F_{2,155} = 3.45; p = .034$) of the three Websites. We do not use the word ‘worst’, because with the ‘least good’ Image map Website, the children evaluated the goodness of the Image map Websites with an 8.2 on a scale from 1 to 10, which is still a very high score (see Table 8). There were no significant differences between the children’s perceptions of hedonic quality (i.e. unruly-manageable) and between their evaluations of fun and beauty of the three Websites.

**What is the Effect of Playful Interface Design on Physiological Measurements of Emotional Arousal**

We also used a more objective method to measure children’s feelings towards the three Website types by measuring electrodermal activity (EDA). From the EDA-data, we computed the number of Electrodermal Responses (EDR) per minute and the total amplitude of the EDR per minute for the period of watching the film and the period of the actual task performance on the experimental Website. Watching the film served as a benchmark, because this was the same over all three conditions. To measure whether there is a difference in EDA between conditions, we computed the difference in number of EDR per minute between the task performance and watching the film for the three conditions. Unfortunately, the EDR data (both number of EDR per minute during the task performance and while watching the film, and the difference between these two variables) were not normally distributed. Therefore, we could not use a parametric test to compute the difference in EDA between conditions and we were constrained to using a non-parametric test.

First of all, we did not measure any EDA during the actual task performance with 47 children. We did not measure any EDA with 64 children while they watched the film. Apparently, for many children, this type of scholarly tasks on a computer did not activate any electrodermal activity at all. Second, using a Wilcoxon signed rank test, there was no significant difference in EDA between task performance and watching the film.

According to an independent samples Kruskal-Wallis test, there was no significant difference in EDA between the three conditions. There was, however, a significant order effect in our study. According to an independent samples Kruskal-Wallis test, there was a significant difference in EDA between children that watched the film before or after performing the actual search tasks. Children that watched the film after the search task performance showed a higher number of EDR per minute during the task performance than children that watched the film before the search task performance. A regular univariate analysis of variance also showed no significant differences in EDA between the three conditions and it also did show a significant order-effect of watching the film before or after the search tasks ($F(1,135)=23.03, p < .05$). Apparently, children were less aroused by the use of the search interface when they first saw the film, than when they saw the film afterwards.

**What is the Effect of Playful Interface Design on Verbal Emotional Utterances?**

Of the children that used the Image map Website, 25 video recordings were studied to collect the verbal emotional utterances about the pragmatic and hedonic quality of the Website. In total, 58 utterances were made. Of these utterances, 55 were related to pragmatic quality and 3 utterances were related to hedonic quality of the Image map Website (see Table 9).

As can be seen in Table 9, most verbal utterances were negative. These utterances were mostly related to the difficulty and inaccessibility of the Website for finding relevant information,
Table 9. Number of positive and negative verbal utterances related to pragmatic and hedonic quality of the Image map Website

<table>
<thead>
<tr>
<th>Quality</th>
<th>Number of Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Pragmatic</td>
<td>7 (12.1%)</td>
</tr>
<tr>
<td>Hedonic</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

such as “I can’t find it at all” or “I find it difficult to search on”. There are only a few positive utterances, such as “This is not really difficult or anything” and “I found it quite fast”. Verbal hedonic utterances are very rare and are more related to pictures on the Website than to the design of the Website itself, such as a child that says: “What an ugly man” (about Columbus). Apparently, children tend to say more about their search activities than about the design of the Website and verbal utterances are almost exclusively related to their perception of the Websites’ pragmatic quality.

Relation between Factors

What is the Relation between Children’s Performance on the Websites and their Attitude Towards these Websites?

To answer this question, we tested the difference for children’s affective responses between tasks that were completed successfully and tasks that were not completed successfully. There was a significant difference between the successful and unsuccessful tasks for children’s ratings of emotional valence ($F = 5.07; df = 1; p = .025$). Children expressed more positive feelings for tasks conducted on all three Website versions when the tasks were completed successfully than when the tasks were completed unsuccessfully (see Figure 9). There was no significant interaction effect between the Website version and whether the tasks were completed successfully or not for the ratings of emotional valence. Furthermore, there were no significant differences for children’s product evaluations (beauty, goodness and fun) between tasks that were completed successfully and tasks that were not completed successfully.

Figure 9. Estimated marginal means of valence (0 = negative feeling; 5 = positive feeling) plotted for the three Website versions with successful and unsuccessful tasks presented on separate lines
Further Diagnosis of Children’s Search Behavior

Searching vs. Browsing in our Experiment

It is not clear from previous research results whether children in general use more searching or browsing strategies. However, most of the results suggest that children prefer to use search engines, but that they are more successful with browsing categories. Even recently, Meloncon et al. (2010) recommended not to include a search engine on a children’s Website, because children have not yet fully developed the intellectual ability necessary to generate relevant search terms (Druin et al., 2009). In addition, giving children the option to search would undermine the process of having them read through the information and explore the Website (Meloncon et al., 2010). In other words, the search engine is seen as a distraction on a children’s Website.

In our data, searching with the search engine scores better than browsing the main categories for almost all components of UX that we have measured. Children were more successful in finding the right content page when they had used the search engine. They also needed significantly less time and fewer clicks to find the information when they had used the search engine than when they had browsed the main categories.

These results are in contrast to results from previous studies on children’s search behavior (Borgman et al., 1995; Schacter et al., 1998; Bilal, 2000). This difference might be caused by a better functioning search engine in our study, although we think that is hard to believe, because the search engine in our study did not work as well as Google, for example. The search engine in our study did not provide query suggestions or spelling corrections while typing a search query. Also, most natural language queries could not be processed by the search engine.

In our experiment, initially 115 children were offered a Website version with a search engine and 43 children without a search engine (see Table 10). This inequality was caused by the fact that many children that were offered a search engine, did not use the search engine at all. We offered children a version with a search engine as often as was necessary, to reach an equal number of children that used the search engine and children that browsed the main categories. We kept offering a Website version with a search engine until the distribution between use and non-use of the search engine was equal over the three Website versions (see Table 10). In total, 77 of the 115 children that were offered a search engine (67%), used the search engine for at least one of the five search tasks. The children that used a search engine, did not automatically use the search engine for all five tasks. Therefore, the percentage of search engine use per task that was conducted on a Website version with a search engine is lower. In total, 294 tasks were conducted by using a search engine, which is 51% of all tasks that were conducted on a Website version with a search engine.

As can be seen in Table 10, the search engine on the Image map was the least inviting to use for the children, because 25 of the 50 children that were offered a search engine on the Image map...
Website, did not use it at all. The non-conventional visual design of the Image map version might have distracted the children’s attention from the search engine as can be seen in Figure 3.

Another explanation for the low percentage of search engine use when a search engine was offered, can be children’s infrequent use of search engines. We asked children how much time they spend using the Internet and most children used the internet less than one hour on both week days and weekend days. When we asked what type of activities they conducted on the Internet, the activities can be ranked from most to least frequently conducted by most children as following:

1. Hyves (a Dutch social network).
2. Playing games.
3. Listening to music/watching video clips.
5. Searching information for myself.
6. Searching information for school work.

Table 10. Number of children that were offered a search engine and number of children that used the search engine

<table>
<thead>
<tr>
<th>Website Version (N=158)</th>
<th>Presence of the Search Engine</th>
<th>Use of the Search Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present:</td>
<td>Used: Not used:</td>
</tr>
<tr>
<td>Classic version</td>
<td>32</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Not present:</td>
<td>6</td>
</tr>
<tr>
<td>Classical play version</td>
<td>33</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Not present:</td>
<td>7</td>
</tr>
<tr>
<td>Image map version</td>
<td>50</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Not present:</td>
<td>25</td>
</tr>
</tbody>
</table>

Internet activities for which a search engine is required (by searching information for themselves or for school work) are conducted least frequently according to children’s self-reports. The children probably do not use a search engine for the Internet activities that are conducted most frequently, such as using Hyves or playing games. The fact that children do not use search engines frequently is a plausible cause for the fact that many children did not use a search engine in our study, even though a search engine was provided in many of these cases.

Search Engine Strategies

In this section, we will analyze children’s strategies and skills using a search engine. What makes keyword searching in our experiment successful? Do children experience the same problems or not with key word searching as reported in previous research? We will describe whether children are able to formulate a query and to select a search result that is coherent with the search task. For this qualitative study of submitted queries and selected search results, we have analyzed the submitted queries of the children that used the search engine on two of the three Website types: the Classical play Website (26 children) and the Image map Website (25 children).

For the qualitative data analysis of search engine use, we analyzed 190 search tasks conducted with the search engine. These 190 search tasks consisted of 322 search attempts to formulate a query. Most search tasks (N = 117) consisted of one query attempt, but some children required more attempts to formulate a search query within one search task (40 * 2 attempts, 20 * 3 attempts, 7 * 4 attempts, 3 * 5 attempts, 2 * 6 attempts and 1 * 10 attempts).

Quality of Query Attempts: Most queries consisted of one word (117 query attempts) or more than one word (165 query attempts) and 40 queries existed of a whole sentence. Natural
language querying is not as frequently applied by the children in our study as reported in previous studies (Marchionini, 1989; Druin, et al., 2009). Most query attempts did not require any help with query formulation or spelling help from the test instructor (87% of all query attempts). The children had less problems with query formulation and spelling than expected from the literature (Borgman et al., 1995; Hutchinson, 2005). This was also shown by the fact that 79.5% of the query attempts were spelled correctly. An important reason was that the children in our study received the search tasks written on a task sheet, so they could use the correct spelling of words from the written task sheet. However, still spelling mistakes were made, although children could read the correct spelling from the task sheets.

Correct spelling does not automatically lead to finding the right content page, because almost half of the correctly spelled queries (44.5%) did not lead to a successful search result. Of course, the success of a search query also depends on the relevancy of the search query and the quality of the search engine.

One way to evaluate the quality of the search queries is to compute their latent semantic analysis (LSA) scores in comparison to the search task, following the example of Kitajima, Blackmon, and Polson (2000) that used LSA to simulate Web navigation. LSA is a technique in natural language processing, in particular in vectorial semantics, of analyzing semantic coherence between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text. Values close to 1 represent high semantic coherence while values close to 0 represent low semantic coherence. For the calculation of LSA-scores, spelling mistakes were corrected. We did not find an effect of high LSA-scores on task success. The mean LSA-score of the successful search queries was .26 and the mean LSA-score of the unsuccessful search queries was .24. LSA-scores do not predict the quality of search queries in our study very well. We will illustrate this with the task about Columbus (see Table 11). The query ‘Columbus’ had the highest LSA-score, but did not lead to success, because the answer to the search task was not mentioned on the content page about Christophorus Columbus, but on the content pages about ‘Discoveries’ or about ‘Vikings’.

**Quality of Search Results’ Selection:** When we look at the behavior of the children in selecting search results, we first of all see that most children tended to directly choose the first search result from the list. From the search attempts in which children actually selected a search result (N=216), 110 selected search results were the first result from the results list. Further, almost all selected search results (211) were selected from the first 10 results that were presented in the results list. This tendency to select the first search result or a search result from the first 10 search results presented was also reported in previous studies (Bilal, 2000; Druin et al., 2009).

The children were not inclined to select more than one search result from the search results that were presented for one search query. Most children only selected one search result (86%)

<table>
<thead>
<tr>
<th>Search Task</th>
<th>Frequent Search Queries (N)</th>
<th>LSA-Score</th>
<th>Successful?</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;It is often said that Christopher Columbus discovered America. Some say that this is not true. They say that another nation discovered America 500 years before Columbus did. Can you find out what people discovered America before Columbus did at Junior Winkler Prins online?&quot;</td>
<td>- America (19)</td>
<td>0.27</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>- Discovery America (17)</td>
<td>0.27</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>- Columbus (10)</td>
<td>0.30</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>- Christopher Columbus (7)</td>
<td>0.33</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>- Vikings (3)</td>
<td>0.13</td>
<td>Yes</td>
</tr>
</tbody>
</table>
of the search attempts). Two search results were selected within 11% of the search attempts and more than two search results were selected within 3% of the search attempts.

Of all analyzed search tasks that were conducted with the search engine (N=190), 83% (after one or more unsuccessful search attempts) were successful and led to the right content page. Within 96% of the selected search results no help was provided by the test instructor. From our analysis it is not entirely clear what the main reason is for the unsuccessful search tasks. A possible explanation is the fact that often children chose high-scent incorrect links. However, our research was not concerned with the quality of the search engine and therefore we did not further analyze the relevancy of the provided search results by the search engine.

Navigation and Browsing Categories

In this section, we will analyze children’s navigation behavior. What makes browsing successful or unsuccessful? Do children experience the same problems or not with browsing as reported in previous research? For this qualitative study of selected main and subcategories and content pages, the navigation paths for the five search tasks of 46 children were analyzed that were either successful or unsuccessful in completing the search tasks by browsing the categories. We will describe problems with the layout of the Websites, problems with the information structure of the Websites and operational navigation problems. The problems presented are not an exhaustive list of all problems, but give an illustration of the most important navigation problems.

Problems with the Location of the Search Box: The most important problem in our view with the layout of the Websites, and particularly with the Image map Website, is the location of the search box. Many children did not use the search engine when it was provided. We do not know for sure if these children were not familiar with using a search engine or whether they did not notice the search box at all. The unconventional layout of the Image map Website and the unconventional location of the search box in the top left corner may have caused the search engine to have been unnoticed by many children (see Figure 10). In a corpus study of children’s informational Websites Jochmann-Mannak et al. (2012) reported that 43% of the Websites with a search engine, presented the search engine in the top right corner of the page and 24% presented the search box in the top center of the page. Therefore, the top left corner is a non-conventional location for children’s informational Websites. However, on the other two versions, the search box was also placed on the top left side, but on these two versions more children used the search engine when offered than on the Image map version.

Problems with the Layout: Most problems with the layout were experienced with the Image map Website, most likely because of the unconventional interface design. When parsing the homepage, children directly focused on a particular part (mostly on the main menu) of the Classic Website and the Classical play Website. This was in contrast to children's parsing behavior on the Image map. Most children parsed the entire homepage of the Image map and all main category links were looked at extensively. This was caused by the fact that the main menu covers the entire homepage of the Image map. There were many problems with the main category links on the Image map Website, because children had to ‘mine sweep’ the images to see the verbal link labels. However, many children did not directly recognize the main category links as such and did not notice their click ability, as can be seen in Figure 10.

Problems with the Information Structure: Problems with selecting the right main and subcategories occurred at all three type of Websites, but mostly at the Classic and the Image map Website. Especially choosing the right main category was problematic for many children, as was also reported
How Interface Design and Search Strategy Influence Children’s Search Performance and Evaluation

by Hutchinson et al. (2006). Often, children had wrong expectations of the content behind main category links. For example, the main category link label ‘human’ was often selected for tasks about the first astronaut or about Columbus, because humans were mentioned in the task description. However, in fact the main category ‘human’ was about the human body.

Children did benefit from the addition of images with the main and subcategory link labels at the Classical play Website in comparison to the Classic Website (see Figures 11 and 12). The images helped children to choose the right main and sub categories and the children used more trial-and-error in choosing the right categories at the Classic Website.

The addition of images at the Image map Website often had an opposite effect on children’s navigation behavior, because the images without the verbal link labels often did not speak for themselves (see Figure 10). In other words, addition of images to the category labels helps children in navigation, but corresponding text labels with the images are essential for quick understanding of the link labels.

Operational Navigation Problems: Children needed more assistance in operating the navigation tools on the Image map than on the other Website versions. The sub category link labels on the Image map appeared in a pop up layer across the homepage and often children did not know how to get out of this pop up layer (see Figure 13). There were no navigation tools provided to go back to the homepage, as was the case at the other two Website versions. Children had to find out for themselves that they had to click somewhere on the screen next to the popup layer to make the popup layer disappear.

Children’s Navigation Skills: Many problems that children experienced on the Websites cannot be attributed to the design characteristics of the Websites, but are caused by children’s own navigation and information skills. Differences between children’s navigation and information skills are more determinant for differences in search performance than differences between Web design characteristics. We will illustrate this by describing differences between search behavior of children that were successful and unsuccessful in finding the right information on the Websites.
Where successful children showed a clear pattern in link selection, less successful children lacked such a clear pattern. Less successful children especially had problems with the interpretation of the main category link labels. They often selected high-scent incorrect links and they almost never selected low-scent correct links. They often lacked domain knowledge to select a relevant category link or did not keep the initial search task in mind. Also, less successful children often did not recognize when they selected the wrong main category link and got lost in the Website’s menu structure.

Successful children did keep the initial search task in mind and had less problems with interpretation of the main category link labels. They were well aware when they had selected a wrong main category. They were often able to select low scent correct links and followed the optimum navigation path more often than unsuccessful children.

**Browsing Strategies:** Successful children used both a trial-and-error strategy and a ‘think-than-act’ strategy on the Classic Website. On the Classical play Website, successful children only used one strategy: the ‘think-then-act’ strategy. This might be explained by the fact that the link
labels on the Classical play Website were presented with a picture that visually represented the textual link label, allowing children to make a better interpretation of the meaning of the link labels. This made the ‘trial-and-error’ strategy less necessary.

Less successful children exclusively used a trial-and-error strategy to find information on all three Website versions. They often only used one strategy and had no plan B, when plan A did not lead them to the right content page. They had trouble keeping the search task in mind and often applied a ‘loopy’ navigation style, which means that they selected the same incorrect main and sub category links again and again. This looping navigation behavior was also reported by Bilal (2000).

Another problem that occurred with some children while using the Image map Website was that they were hesitant to click on the main and sub category links. The playful lay out of the Image map Website made them insecure to click on the links.

**Processing the Content Page:** Successful children often asked for assistance in understanding the content. Less successful children did not ask for help, but just gave a wrong answer. Reasons for giving wrong answers were because of a wrong interpretation of the text, not recognizing internal hyperlinks to a relevant content page, because of low literacy skills, scanning the text too quickly or by not keeping the initial search task in mind. Successful children mostly had the capacity to reflect on the initial search task and made sure that they kept the search task in mind.

**Processing Problems:** It was noted that children experienced more problems on all areas of navigation on the Image map than on the other Website versions. Children even had trouble understanding the search tasks correctly when using the Image map Website, which did not occur with the other Website versions. This might be explained by the ‘cognitive load theory’ (Sweller, Merrienboer & Paas, 1998). Processing the Image map Website might have taken so much cognitive energy from the children, that there was too little cognitive energy left to interpret the search task.

**Game-Experience:** Finally, we found some clues that game-experience influences children’s navigation behavior. Children with little game-experience (mostly girls according to the results of the profile survey) often think first before they act. Children with a lot of game-experience (mostly boys) use the trial-and-error search strategy. This
trail-and-error strategy is provoked when children do not know exactly where to go, which is often the case on the Image map Website. Therefore, the use of the Image map seems more suited for children with a lot of game-experience.

CONCLUSION AND DISCUSSION

Validation of the Used Methods

In this section we will answer our research questions. The first question is about the suitability of existing methods for research with children to measure feelings and perceptions of pragmatic and hedonic quality that are used in research with adults. We experienced that methods such as the SAM-scale for emotional arousal and the AttracDiff 2 questionnaire are not valid methods in research with children. Unlike Greenbaum et al. (1990), we could not validate the use of the SAM-scale for emotional arousal with children. These opposite results might be caused by the different settings in which the method was used. The pictures in the SAM-scale for emotional arousal might be related easier to ‘fear of the dentist’ by children (which was the case in the research of Greenbaum et al., 1990) than the feeling experienced with using a Website. We could, however, validate the SAM-scale for emotional valence with children.

Almost all picture anchors for the semantic differential scales for the pragmatic and hedonic items could not be validated for use with children in our experiment (which we did not expect based on our pilot test of the semantic differential scales with 14 children). The children in the experiment did not interpret the picture anchors as intended by the developers of the AttracDiff 2 questionnaire (Hassenzahl et al., 2003). The children only made a clear distinction between which end of the scale was intended as positive and which end of the scale as negative in their opinion. The children were not able to make a distinction between perceptions of hedonic and pragmatic quality of the Websites. The pragmatic and hedonic items were all seen as the same construct of valence (i.e. positive versus negative).

Although we did find differences in scores for the product evaluations beauty, goodness and fun, we doubt whether children were able to make a clear distinction between these concepts, as was also reported by Read et al. (2002). Fortunately, we were able to use an objective method to measure emotional arousal by using the Q-Sensor. This method turned out to be valid to measure emotional arousal with children in contrast to the subjective methods used in our experiment.

Based on this experiment, we cannot judge whether the UX models presented in research with adults (Hassenzahl, 2004; Thüring & Mahlke, 2007; van Schaik & Ling, 2008) can also be applied to children’s UX. Although we could not entirely reproduce UX research concerning perceptions of pragmatic and hedonic quality in our experiment, we did show the difficulty of reproducing UX research with children. More research is needed to develop valid methods to measure children’s perceptions of pragmatic and hedonic quality.

Effects of Different Design Types

We did not find any effects on children’s search performance on informational Websites of the different design types. The variation in design of the Websites did not have an effect on search success and efficiency. As opposed to the performance scores, there were differences in the subjective scores measured. Children were most positive about the Classical play Website according to their scores on the SAM-scales for valence. This positive score is most likely not based on pragmatic issues, because both children that were successful and unsuccessful in finding the right information on the Classical play Website, gave this Website higher scores than the other two Website versions. Apparently, their feelings about the Website is based on more hedonic issues, such as whether
they are attracted to the interface design of the Website. Surprisingly, we saw the same pattern for children’s evaluations of goodness: both successful and unsuccessful children evaluated the Image map version as less good than the other two Website versions. There were no significant differences between children that were successful and children that were unsuccessful in finding the right information. Apparently, search success was not determinant for children’s evaluation of goodness.

The Relation between Performance and Attitude

We did not find evidence for a strong relation between affectivity and usability in our study. Children’s affective responses are not based on the effectiveness of search performance on the Websites. However, their affective responses could be based on pragmatic issues besides the final success in finding the information. For instance, their affective responses could be based on the ease of use while interacting with the interface, as was reported by Sim et al. (2006) who found that children appeared to have less fun when their interactions had more usability problems. However, in our study children’s affective responses towards the search systems seem to be independent from their actual search behavior and most likely based on perceived hedonic quality and aesthetics of the interface. This could be best tested by a pre and post measure of affective responses, to see whether actual behavior changes children’s attitude towards product evaluations such as beauty, goodness and fun (van Schaik & Ling, 2008). Because of time constraints it was not feasible in our experiment to conduct both pre and post tests to measure affective responses.

We did find some proof that the children’s affective responses were based on the ease of use while interacting with the interface (i.e. pragmatic quality). An analysis of verbal utterances of 25 children that used the Image map Website, showed that most of the utterances were negative and related to the Websites’ low pragmatic quality. This result supports the fact that the children’s valence scores and evaluation scores for goodness were lowest for the Image map Website and that these scores are based on perception of pragmatic quality. The fact that almost no hedonic utterances were made, supports the fact that playfulness and expressive aesthetics do not have a large influence on children’s attitude towards and evaluation of the Website.

The fact that we did not find a relation between beauty and usability in our experiment can also be caused by children’s tendency to indicate the highest score on the scale (Markopoulos et al., 2008). Most children felt very positive about the three Websites and little frustration was uttered. Therefore, we did not find evidence that pragmatic frustrations lower children’s perception of aesthetics as Tuch et al. (2012) found in their research with adults. This is most likely also caused by children’s tendency to give socially desirable answers. Although they were often not successful at all in conducting a search task, children found it difficult to be negative about the system and tended to blame themselves instead of the system. This was also reported by Serenko (2007), who studied the self-serving biases of interface agent users. He found that adult users may attribute their success to an interface agent and hold themselves responsible for task failure, just like the children may have done in our experiment.

Interaction with the Chosen Search Strategy

From this experiment we can conclude that the search strategy that is used by children is much more determinant for their search performance than the interface design of the search environment. Searching with a search engine proved to be much more effective and efficient than browsing the navigation structure. This was the case for all three types of Websites in our experiment.
We found that there was a significant difference in success scores when children received help from the test instructor. However, this difference meant that children were less successful. In other words, children did not become successful because of the help they received. These findings prove that help was only offered to motivate and reassure children during the search process (as recommended by Markopoulos et al., 2008) and offered help did not have a significant effect on children’s success in finding the right information.

It turned out that searching with the search engine is more effective and efficient than browsing the categories. This is quite logical when considering the fact that children nowadays are members of the ‘Google-generation’ (Rowlands, 2008). However, we did not look at the long-term effects of the fact that children prefer and are better in searching than in browsing. Searching instead of browsing might, for example, have a negative effect on children’s knowledge of information taxonomies. The search engine can be compared to a ‘black box’ that does not give insight in how information is related in a taxonomy. Future research is needed to study long-term effects on children’s knowledge of information architectures of searching with a search engine in comparison to browsing categories.

Search and navigation behavior on the Web is constantly changing. Currently, traditional search engines as Google are loosing ‘traffic’, because people more and more use social media such as Facebook and Twitter as their primary Web entrance to find information (Xiang and Gretzel, 2010). This might also cause changes in children’s search and navigation behavior.

Children’s Navigation and Information Skills

When studying children’s search performance in more detail, we found that many problems that children experienced on the Websites cannot be attributed to the design characteristics of the Websites, but are caused by children’s own lack of navigation and information skills. In other words, children’s navigation and information skills are a better predictor of children’s search success than the type of design characteristics of the search interface. Children with a lot of internet experience, encounter fewer problems with searching and browsing than children with little internet experience. Also, children with a lot of domain knowledge are more successful in formulating relevant search queries, selecting relevant search results, or selecting relevant main and subcategories. Children that are able to check and monitor their own activities, termed metacognition (i.e. the voluntary control an individual has over his own cognitive processes) (Brown & DeLoache, 1978) are more capable of keeping the initial search task in mind, or of recognizing when they select the wrong main or sub category link.

However, particular search characteristics might support particular groups of children. For example, our research provides indications that children with a lot of game experience are better off with the Image map Website, because this interface type is based on exploring and a trial-and-error strategy. Our research also provides clues that children with little internet experience are better off with the Classic Website, because there are not so many visual stimuli to distract these children from their navigation path.

An important lesson learned from this experiment is that the variance between children is much more determinant for differences in search success and performance than the variance between search interfaces. This makes the challenge for designers to design interfaces that support children in effective information-seeking even greater. It also stresses the importance of educating children in navigation and information skills. However, the most important lesson learned from this experiment is the fact that children’s search success and search efficiency is much larger when they use a search engine, than when they browse the main categories of the Website.
ACKNOWLEDGMENT

We would like to thank the children that cooperated in this study. We would also like to thank Matthijs Noordzij of the University of Twente for his help processing the data of the Q-sensor and MA students, Jiske Naber, Marjolein Makkings, Aafke Ariaans, Yvonne Joosten and Amy Mooij, for their assistance in conducting the qualitative analyses of the experimental data. The research is funded by the Netherlands Institute for Public Libraries (SIOB) and the European Community’s Seventh Framework Programme FP7/2007-2013 under grant agreement no. 231507, named ‘PuppyIR’.

REFERENCES


KEY TERMS AND DEFINITIONS

**Browsing:** Browsing is a search method in which the user selects a relevant subject category from a menu. The user can ‘recognize’ a relevant category from the provided categories in the menu.

**Classic Interface Design:** Classic interface design aims at simplicity, consistency and focus. The layout of classic interfaces is kept minimal with little graphic elements; key elements of the interfaces are the center of attention and different pages have the same layout. Page components are located on conventional locations. The navigation is a basic, textual menu of categories presented in a horizontally or vertically presented list.

**Emotional Arousal:** The extent to which the user gets excited (by interacting with an interface).

**Emotional Valence:** The extent to which the user’s feelings (towards an interface) are positive or negative.

**Keyword Search:** Keyword search is a search method in which a search engine is used to search for information in a vast information space by submitting a search query. The search engine matches this query to information elements in the information space and presents search results from which the user can select a result that is relevant for his search task. The user has to ‘recall’ and formulate a search query from his memory.

**Perception of Hedonic Quality:** The user’s perception of the extent to which interacting with an interface is stimulating for the user and the extent to which the user can identify with the interface (i.e. pleasure-producing quality of an interface).

**Perception of Pragmatic Quality:** The user’s perception of how user-friendly an interface is and how effective and efficient search performance is when using this interface (i.e. user-perceived usability).

**Playful Interface Design:** Interface design can be made playful both in design of navigation as in visual design. With a playful navigation approach, category labels are integrated in a screen filling image, often without textual labels. Users have to ‘explore’ the screen image in search of categories, which makes interacting with the interface a playful experience. Visual design can be made playful by adding many different colors, images and animations to the interface and by playfully arranging visual elements on the screen.

**Search Performance:** The effectiveness and efficiency (for example in amount of time and clicks needed) with which a user can search and find relevant information by interacting with a digital interface.

**User Experience:** The research field of User Experience (UX) argues for a broader perspective for evaluation of systems or interfaces than usability. Besides instrumental qualities (i.e. pragmatic quality/usability), also non-instrumental qualities are important, such as visual aesthetics, hedonic quality or haptic qualities.

ENDNOTES

1. Six of the AttractDiff 2 items were not used in our study, because these verbal anchors were too difficult to translate in a picture that we thought could be well-interpreted by children: isolating – integrating, gaudy – classy, takes me distant from people – brings me closer to people, typical – original, conservative – innovative, commonplace – new.

2. Jiske Naber and Marjolein Makkinga (students of the Master Communication studies) assisted the authors by analyzing verbal utterances of 25 children that conducted search tasks on the Image map Website.

3. Aafke Ariaans (MA student) assisted the authors by analyzing the reliability and validity of the data.

4. Yvonne Joosten (MA student) assisted the authors by conducting qualitative analyses of search engine use.

5. Amy Mooij (MA student) assisted the authors by conducting qualitative analyses of the children’s navigation paths.
APPENDIX 1

“How interface design and search strategy influence children’s search performance and evaluation”

Task Descriptions Translated from the Original Dutch Versions with the Optimum Navigation Path Presented for Each Task

Task 1

You’ve seen a chameleon in the zoo and you saw that he was moving very slowly. Now you wonder how a chameleon captures its prey, because he seems far too lazy for that. Can you find out how a chameleon catches its prey at Junior Winkler Prins online?

Optimum Navigation Path Task 1:

- **Correct Main Category:** Nature
- **Correct Subcategory:** Reptiles and amphibians
- **Correct Content Page:** Chameleon

Task 2

You would like to become an astronaut and explore space in search of alien planets, like in the movies. Can you find out who was actually the first living creature that traveled through space at Junior Winkler Prins online?

Optimum Navigation Path Task 2:

- **Correct Main Category:** Science and technology
- **Correct Subcategory:** Space and Space travel
- **Correct Content Pages:** Space/ Astronaut

Task 3

You spend your holiday in Friesland with your parents and your father told you that he participated in the famous Dutch skating tour named ‘Elfstedentocht’ once. He also told you that he met the Dutch crown prince Willem-Alexander during the tour. Can you find out under what name Willem-Alexander participated in the skating tour that day?

Optimum Navigation Path Task 3:

- **Correct Main Category:** Sports and exercise
- **Correct Subcategory:** Stadiums & Tournaments
- **Correct Content Page:** “Elfstedentocht”
Task 4

It is often said that Christopher Columbus discovered America. Some say that this is not true. They say that another nation discovered America 500 years before Columbus did.

Can you find out what people discovered America before Columbus did at Junior Winkler Prins online?

Optimum Navigation Path Task 4:

- **Correct Main Category**: History
- **Correct Subcategories**: Time of cities and states / Time of discoverers and reformers / Time of monks and knights.
- **Correct Content Pages**: Vikings / Discoveries.

Task 5

Isaac Newton invented a device to be able to see the stars better. Can you find out what the name of that device is at Junior Winkler Prins online?

Optimum Navigation Path Task 5:

- **Correct Main Category**: Famous people / Science and technology.
- **Correct Subcategories**: Inventors and scientists / Space and space travel / Measuring, weighing or counting / How it works.
- **Correct Content Pages**: Isaac Newton / Star / Telescope.
APPENDIX 2

Bipolar Picture Anchors for Pragmatic and Hedonic Quality from the AttracDiff 2 Questionnaire

Pragmatic Quality (PQ)

Figure 14. Technical – human

Figure 15. Complicated – simple

Figure 16. Impractical – practical
Figure 17. Cumbersome – direct

Figure 18. Unpredictable – predictable*

Figure 19. Confusing – clear

Figure 20. Unruly – manageable
Hedonic Quality – Identification (HQI)

Figure 21. Amateurish – professional

Figure 22. Cheap – valuable

Figure 23. Non-inclusive – inclusive*

Figure 24. Unpresentable – presentable
Hedonic Quality – Stimulation (HQS)

* Picture anchors that were left out of the final experiment based on the results of the pilot study.