D5.5 Demonstrator evaluations for novel interaction models

File name	PuppyIR-D5.5-Demonstrator-Evaluations
Author(s)	Betsy van Dijk (UT) Saskia Akkersdijk (UT) Carsten Eickhoff (TUD) Andreas Lingnau (UoS) Frans van der Sluis (UT) Kimberly Snoyl (UT)
Work package/task	WP5
Document status	draft/ready for approval/final
Contractual delivery date	M36
Confidentiality	Public
Keywords	User Evaluation, Demonstrators, Usability
Abstract	In this deliverable we present the results of the evaluations of the Museum Demonstrator and the Hospital Demonstrator.

Table of Contents

Executive Summary

This report presents the results of the user evaluations of the two main demonstrators that have been designed and developed for the PuppyIR project: a Museum Demonstrator and a Hospital Demonstrator.

1 Introduction

Two main demonstrators have been designed and developed for PuppyIR: a Museum Demonstrator and a Hospital Demonstrator. The Museum demonstrator focuses on interaction and interfaces, while the focus of the Hospital Demonstrator is mainly on information processing and presentation. An overview of the design, development, and features of the demonstrators can be found in Deliverable 7.2 (Museum Demonstrator version 2.0) and Deliverable 7.4 (Hospital Demonstrator version 2.0). Details about the evaluation measures used can be found in the report of Deliverable 5.4 (User Evaluation Toolkit) and in the toolkit on the PuppyIR site (see http://www.puppyir.eu/results/user-evaluation-toolkit).This report describes the evaluation of the demonstrators.

The Museum Demonstrator aims to enrich children's experiences during a museum visit. It creates an interactive museum visit using a multi-touch table, terminals with touch input, and marker tracking. Before the demonstrator was developed, PuppyIR partner Museon already used an electronic quest that children could choose to answer during their visit. Therefore terminals with touch input and a bar code scanner are available all over the exhibition area. The admission tickets have a bar code on the back what can be scanned at a terminal. The first time a ticked is used with one of the terminals a quest is generated with random questions from a central database. The visitor is asked to register his or her name to personalise the quest and will then get 12 questions about different topics in the exhibition area. When the admission ticket is scanned at one of the terminals, the next question appears until it is answered. Once the question is answered the visitor gets immediate feedback whether or not the answer is correct before the next question is displayed.

We used this existing infrastructure and enhanced it by a tabletop device that has been placed in the museum's entrance area. It has a multi-touch surface and can identify fiducial markers, unique identifiers, similar to the concept of 2D bar-codes. Each visitor who wants to use the table gets a ticket with both a fiducial marker and a bar-code. The ticket can be used as an identifier on the table and with the terminals in the exhibition space.

For the Museum Demonstrator we developed a collaborative application for the multi-touch table that consists of two parts. The first part, from now on called the initial game, can be used by up to four children simultaneously to browse through the different exhibition topics and to determine the contents of the interactive quest together. The second part, the end game, is used after the tour through the exhibition space to provide further information about the visited exhibits. In the end game children can choose topics/exhibits to create a personalised virtual museum catalogue. The children can use this catalogue for internet search or learning tasks after the visit. It allows them to access information about the topics/exhibits they were most interested in. As part of the PuppyIR project a special website has been created for the post-visit use of these topics. This website uses the PuppyIR framework.

The Hospital Demonstrator, also known as the Emma Search (EmSe) engine, has the goal to improve the accessibility of medical information for children. When undergoing medical treatment, often in combination with extended stays in hospitals, children have been frequently found to develop an interest in their condition and the course of treatment. However, finding information related to medical conditions is often a difficult and sensitive task. Consequently, designing and developing search services for children presents a number of challenges, including: children's problems expressing complex information needs, finding and crucially identifying relevant information, and ensuring that information is understandable, appropriate, and sensitive to the child's physical and emotional state. To address these challenges, we developed the PuppyIR Hospital Demonstrator for the Emma Kinderziekenhuis (EKZ) at the Amsterdam Medical Centre (http://www.emmakids.nl/). EmSe is envisioned to be accessible to staff and patients within the hospital (via bedside and other terminals in the hospital), and also to out-patients via the web. The Hospital Demonstrator has been evaluated in two stages. The first stage is to obtain

feedback from the staff at the hospital. We used the results of this initial evaluation to refine the demonstrator to incorporate suggested changes before the first patient contact. In the second stage, we evaluated the demonstrator with patients.

In sections 2 and 3 we report about the evaluation of the Museum Demonstrator. Section 2 focuses on the evaluation that took part during the Museon visit of two Dutch primary school classes. In section 3 we focus on the evaluation we did at the school, with the same two school classes. In sections 4 and 5 we report about the evaluation of the Hospital Demonstrator. Section 4 reports about stage 1 of that evaluation, the evaluation with the hospital staff. In section 5 we report about stage 2, the evaluation with children in the hospital.

2 Evaluation of the Museum Demonstrator at Museon

Using technology in order to add a social dimension to a museum visit can support and invoke interaction and collaboration between young visitors of a museum (Yiannoutsou et al., 2009). As a main target, we wanted to enrich children's experiences during the museum visit by helping them to collaborate with other visitors on shared interests and by implicitly guiding them through the museum exhibition. One main source of daily visitors are school classes usually visiting the exhibition in the morning, while in the afternoon, at weekends and school holidays there are more individual visitors than groups. The target group of the first pilot study with a prototype touch table application in Museon - which has been described in Deliverable 2.4 - is the general public. The Museum Demonstrator aims at school classes that first visit the museum and later, back at school, continue to work with relevant contents. In this section the evaluation of the Museon part of the Museum Demonstrator will be treated.

2.1 Setup of the experiment

An experiment was conducted in Museon to find an answer to the question if the addition of the collaborative initial game at the multi-touch table, used to determine the contents of the interactive quest together, supported and invoked interaction and collaboration between the participants of the experiment. In addition, we evaluated the educational and fun experience of the entire museum visit.

2.1.1 Experimental conditions

The experimental setup consisted of two conditions. In one condition the children started the museum visit at the multi-touch table, where they selected topics they were interested in by choosing images that represented these topics. The chosen topics were parts of the exhibition and were used to determine a route through the exhibition room of the museum. Based on the results of the initial game, the children starting at the table received a personalized quest of twelve questions to be answered at twelve different exhibits. In the second condition the children started with an electronic quest that was generated with random questions from a central database. Hence in this condition the children did not start with the collaborative initial game at the table and the quest was not personalized.

2.1.2 Participants

The experiment took place at Museon in The Hague, the Netherlands. Two school classes of a Dutch primary school in The Hague (Nutsschool M.M. Boldingh), a school located close to Museon, participated in the study. In total there were 48 children who participated, 21 children aged 11-12 years old that were in their final year (8^{th} grade) of primary education and 27 children of 10-11 years old from the pre-final year (7^{th} grade).

2.1.3 Procedure

The teachers formed groups of four children (sometimes three if necessary) before the museum visit. The children of the 7^{th} grade were divided in seven groups and the children of grade 8 were divided in six groups. In total seven groups (four from grade 7 and three from grade 8, in total 26 children) were assigned to the condition that started at the multi-touch table and six groups (three groups from both grade 7 and grade 8, in total 22 children) were assigned to the condition that did not start at the table.

The groups got an explanation about the procedure of the experiment. They were also told that they could ask for help if something was not clear. Then roughly half of the groups started at the table with the initial game followed by a personalized quest, while the other half immediately started with a quest of twelve questions through the permanent exhibition. After all members of a group had finished the quest, they went to the multi-touch table to get further information about

the visited exhibits and to choose topics/exhibits they were most interested in. All teams did the end game at the table, also the teams that did not start at the table. Their choice was used later, in the experiment that evaluated the school part of the evaluation.

After the end game the children had to hand in their tickets and they filled in a questionnaire that they had to hand in to their teacher afterwards.

2.1.4 Measures

A questionnaire was used to measure the constructs enjoyment and collaboration. Subscales for the constructs were inserted in the questionnaire. The complete questionnaire can be found in the user evaluation toolkit on the PuppyIR site (see http://www.puppyir.eu/results/user-evaluation-toolkit). The questionnaire also contained a few open questions to see what the children learned. In addition datalogs were used to derive a quest score.

Enjoyment and intrinsic motivation

We measured enjoyment in three different ways that each focus on a different aspect of enjoyment. The first measure used is the Smileyometer taken from the fun toolkit for children of Read and Macfarlane (2002).

Figure 1 The Smileyometer

The Smileyometer is based on a 5-point Likert scale and uses five smileys, especially designed for children. See Figure 11 for an example. For each of the three parts of the quest the children used the Smileometer to answer the question “How much fun was it to do that part?” In Deliverable 5.4 (the report that accompanies the User Evaluation Toolkit) we argue that the Smileyometer is a useful tool for children of 10-12 years old.

The second enjoyment related measure we used is the Again - Again table (see Figure 2), also from the fun toolkit. The Again - Again table measures engagement. The children were asked if they would like to do the activities (initial game, quest, end game) again. This measure is based on the knowledge that people like to do fun things again.

Would you like to do it Again?

	Yes	Maybe	No
Museon quest	✓		

Figure 2 The Again-Again table.

Although the results of the Again-Again table are often highly correlated with the results of the Smileyometer, we used both measures because the difference in emphasis of the evaluation (judging the software versus giving your own opinion on what you like to do again) might have influence in some cases. See the report of Deliverable 5.4 (User Evaluation Toolkit) for more details on this decision.

The third measure we used to measure enjoyment in the experiments in Museon was the Children IMI interest/enjoyment scale. This scale was derived from the Intrinsic Motivation Inventory (IMI), a multidimensional measurement device which can be modified to fit specific activities (University of Rochester, retrieved 2012). The interest/enjoyment subscale we used is considered the self-report measure of intrinsic motivation and is developed for use by adults. Although this scale basically also measures fun, like the tools of the Fun Toolkit, this scale measures interest as well. If we manage to get a reliable scale for children here, the constructs enjoyment and interest are coupled to get a measure of intrinsic motivation, a very important factor for learning. We adapted the IMI interest/enjoyment scale to make it more suitable for children. We reversed the negatively formulated statements and used a 5-point scale with 'Totally disagree' at the negative end and 'Totally agree' at the positive end and we used the smileys of the Smileyometer. See Figure 3.

Figure 3 Answer categories of the Children IMI interest/enjoyment scale

The new Children IMI interest/enjoyment scale is part of the User Evaluation Toolkit on the PuppyIR site and more explanation on the development and use of the scale can be found in the report of Deliverable 5.4.

Collaboration

To measure perceived collaboration, we added three items to the questionnaire:

- Ik heb heel erg mijn best gedaan om anderen te helpen bij de speurtocht
  (I tried very hard to support others on doing the quest)
- Ik heb veel samengewerkt met mijn klasgenoten
  (I collaborated much with my classmates)
- Ik vond het leuk om anderen te helpen tijdens de speurtocht
  (I liked supporting others during the quest)

We also observed the children while they did the quest but it was very difficult to keep track of the many things that happened because many groups did the quest almost simultaneously. Hence no systematic observation results could be obtained.

Education

To test whether the children had learned from doing the quest the questionnaire contained four open questions about a dinosaur they had seen in the museum. To make sure that everybody saw the dinosaur, the children that started at the multi-touch table were asked to add the dinosaur to their collection. The children who had a random pre-prepared route, all had the dinosaur included in their route (hence it was not completely random for the sake of the experiment). For each of these four questions 1 point could be earned. If the question was not answered completely wrong but also not completely right 0.5 points were given, unanswered or wrongly answered questions got 0 points.

The quest score (a score between 0 and 100, where the score of 100 could only be achieved if all the questions were answered correctly at the first attempt) was taken from datalogs of the quest in a Museon database.

2.2 Results

2.2.1 Enjoyment and intrinsic motivation

Table 1 presents the results for Enjoyment for the three main parts (initial game, quest, end game), as measured with the Smileyometer on a 5-point Likert scale.

	Table at start	**No table at start**
Enjoyment initial game	Mean=3.77, SD=0.77	-
Enjoyment quest	Mean=3.50, SD=0.95	Mean=3.00, SD=1.27
Enjoyment end game	Mean=3.54, SD=0.86	Mean=3.95, SD=1.09

Table 1 Results on enjoyment for the three main parts.

Using the Mann-Whitney test no significant differences were found between the two experimental groups on enjoyment of the quest and the end game. We found significant differences between age groups: children of grade 7 scored higher than children of grade 8 on enjoyment of the quest and initial game. More details about these age differences can be found in Deliverable 5.4.

The Again - Again table was used in three questions about the initial game, quest and end game. Users could choose between no, maybe and yes on the question if they want to do an activity again. The results are shown in Table 2. Chi-square tests show that there was no significant difference between the two conditions for the quest part of the visit. For the end game there was a significant difference: children who already used the multi-touch table at the start of the visit were more positive to use the table again than children who used the table in the end game for the first time (chi-square=6.06, *df*=2, *p*=0.048).

		Table at start	**No Table at start**	**Total**
Again - Again table initial game	No	0	-	
	Maybe	10	-	
	Yes	16	-	
Again - Again table quest	No	5	6	11
	Maybe	9	7	16
	Yes	12	8	20
Again - Again table end game	No	2	5	7
	Maybe	7	10	17
	Yes	16	6	22

Table 2 Enjoyment measured with the Again - Again table for the three main parts.

Enjoyment/Intrinsic motivation was measured by the Children IMI interest/enjoyment scale, an adapted version of the Intrinsic Motivation Inventory subscale. This adapted scale appeared to be very reliable for the children that participated in our study (Cronbach alpha= 0.923). Hence we used the mean score on this enjoyment/motivation scale in our analyses.

The mean score on this Enjoyment scale was 3.54 (SD=1.01) on a 5-point scale which is quite high but not extremely high and variability seems to be high enough, indicating that the scale is useful for children of 10-12 years old. We used the Mann-Whitney test to check if there was a significant difference between the scores of the two experimental conditions on this enjoyment/motivation scale. No significant difference was found. We did find significant age differences between the results of the class from grade 7 (10-11 years old) and the class from grade 8 (11-12 years old). The mean enjoyment/motivation score of grade 7 was 3.96 (SD=0.88) and the mean score of grade 8 was 3.003 (SD=0.94). This mean score was significantly higher for children from grade 7 than for children from grade 8 (Mann-Whitney U=123.5, *p*=0.001).

2.2.2 Collaboration

We tested the three items of the questionnaire that measure perceived collaboration on reliability. The Cronbach's alpha for the three questions was 0.649 which is not really high enough (we use a threshold of 0.7) . If we delete the question “I collaborated much with my classmates” we get a 2 item scale with a Cronbach's alpha of 0.763. Hence we analyse this scale about supporting others and the question about perceived collaboration separately.

The mean score on the supporting others scale was 3.73 (SD=0.87) for the group that started at the table and 3.52 (SD=1.13) for the children that did not start at the table. According to the Mann-Whitney test there is no significant difference between the two experimental groups. For perceived collaboration the mean scores of the experimental groups were 3.77 (SD=0.95) for the table group and 3.55 (SD=1.10) for the group with no table at the start. Again the difference in means is not significant. Using the Mann-Whitney test we did not find any significant differences between the children of grade 7 and the children of grade 8 on perceived collaboration and on the supporting others scale.

2.2.3 Education

The mean score on the open dinosaur questions in the questionnaire was 1.57 (SD=1.11) for the children that did not start at the table and 1.75 (SD=0.95) for the children that started at the table. According to an independent t-test this difference in means is not significant. We did not find any differences between the results of grade 7 and grade 8 either.

The mean scores on the quest were 75.38 (SD=16.71) for the table group and 72.82 (SD=9.3) for the group that did not start with the table. Again no significant differences were found between the experimental groups. Again, no significant differences in the mean scores were found between the different age groups. Notice the high standard deviations. Because in some groups children cooperated and filled in the answers on only one ticket (resulting in a very high score for one child and a low score for the others of the group) while in other groups some children did not finish the quest at all (and hence got a low score) these scores are not very reliable.

2.3 Discussion and conclusion

No significant differences were found between the two experimental groups on enjoyment as measured by the Smileyometer and the Children IMI interest/enjoyment scale. Moreover we did not find any significant differences between the experimental groups on collaboration and on how much they learned. We did find one significant difference between the conditions: On the Again – Again table measure there was a significant difference for the end game: children who already used the multi-touch table at the start of the visit were more positive to use the table again than children who did not start at the table. We do not really have a good explanation for that.

Getting back to our main research question: we did not find clear evidence that the collaborative initial game at the multi-touch table supported and invoked interaction and collaboration between the participants of the experiment and enriched the children’s experience during the museum visit. The results on the enjoyment scales were quite high and where we found small differences they were in favour of the condition with the multi-touch table at the start but there is not enough evidence to prove the added value of the multi-touch table on the experiences during the quest. Of course as an extra attraction, the table will have a value of its own.

We did find significant differences between age groups: children of grade 7 scored higher than children of grade 8 on enjoyment of the quest and initial game as measured by the Smileyometer. In addition the overall mean score on the Children IMI interest/enjoyment scale was significantly higher for children from grade 7 than for children from grade 8.

Probably the most important result of this evaluation is the fact that we adapted the IMI scale for adults to get a reliable scale to measure intrinsic motivation of children of 10-12 years old.

3 Evaluation of the Museum Demonstrator at School

The two primary school classes that visited Museon (see section 2 of this report) participated in the school evaluation of the demonstrator as well. A few weeks after the children had been to Museon we visited the school to evaluate the school-part of the demonstrator.

3.1 Setup of the experiment

The experiment was conducted in the school as part of a normal school day. We used the computer facilities at the school, which consisted of four computers in the corridor next to each classroom. We used the computers of the two classes that participated and were allowed to use the computers of two additional classes that were close to the classrooms of our participants.

3.1.1 Participants

The participants in this study were 46 children from a Dutch primary school in The Hague: one group from the final year (8th grade) of primary education, existing of 19 children of 11-12 years old and one group from the pre-final year (7th grade) existing of 27 children of 10-11 years old. Due to problems saving the screen recordings, two video's that could not be annotated, and one child that did not take the tasks seriously, a total of 40 children were included in this study, 23 children of the 7th grade and 17 children of the 8th grade. All children but one participated in the visit to the Museon described in section 2.

3.1.2 Materials

The website we used (see Figure 4) has a image-based interface. The upper part of the interface consists of a topic menu. The categorization of the menu is based on the categories Museon uses in the permanent exhibition: zon, zee, strijd, kunst, bot, steen, mens and religie. We added the category `mijn selection' (my selection) that contains the topics that children selected during their museum visit (their personal virtual museum catalogue). By selecting one of the categories its topics are displayed in the horizontal bar below the categories. The images used here are images provided by Museon which the children also used at the multi-touch table in Museon.

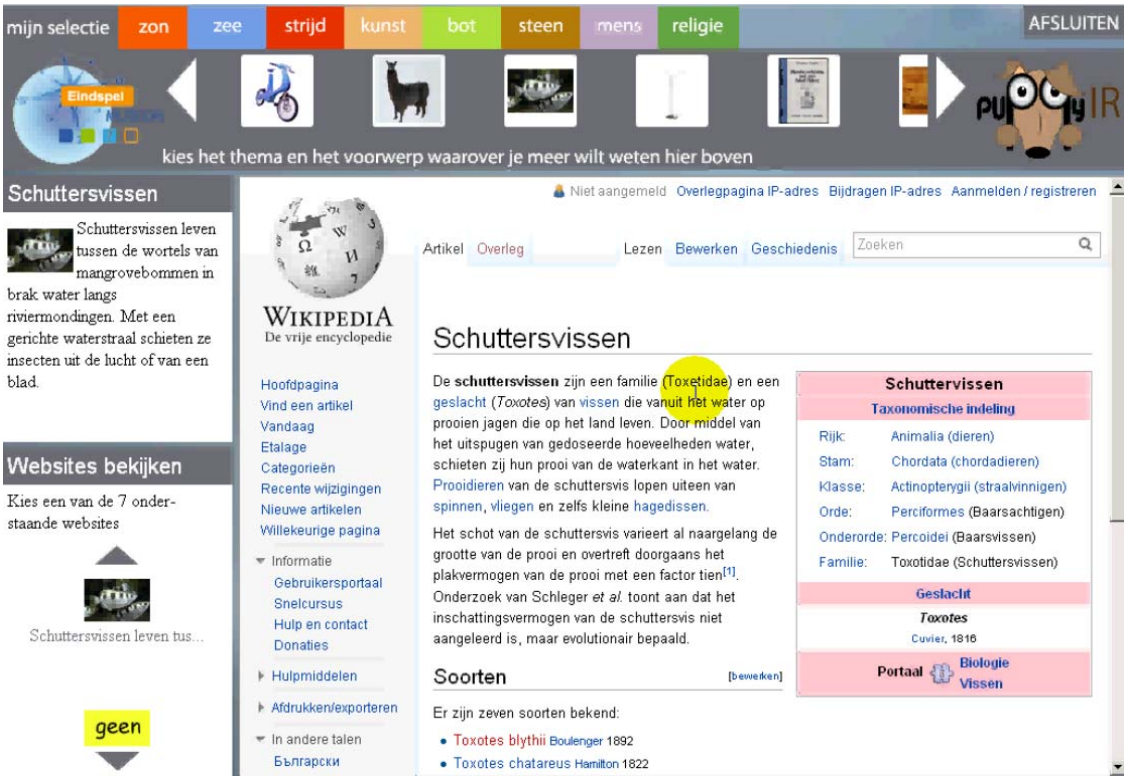

Figure 4: Screenshot of the used website: upper part - topic menu (categories and topics of the selected category); left side top part - general information of the selected topic; left side bottom part - results of an internet search of the selected topic; main window/right bottom part - content of the selected internet search result.

When one of the topics is selected two things change on the left side of the interface. The top part displays some general information about the selected subject and the bottom part displays results of an internet search on that subject. These results are presented using images from the retrieved websites. By clicking one of the internet search results, the content will be displayed in the main window (right bottom part) of the interface (see Figure 4).

The children used this interface for three different types of search tasks. Two of the tasks were fact-based, one was an open-ended task. Each child was asked one question of each type, depending on the topics he/she chose during the Museon visit. The first and second task required the children to make use of different parts of the website. The first task was a fact-based question. For answering this question children could make use of the 'my selection' part of the website which is displayed first after entering the website. The second task again was a fact-based question but now it was not possible to use 'my selection' because we chose a subject that none of the children chose during the Museon visit. Therefore the children were forced to use the menu structure. This way we would get a better impression of how the children navigate. In the third task – an open-ended task where children were asked to find information about a certain topic and make a short summary of the things they found - they were able to use the 'my selection' part of the website again for finding the right topic.

Screen recordings were used to collect information about the search behaviour of the children.

3.1.3 Procedure

Children were retrieved from their classrooms for the experiment after they were informed by their teacher that they were going to participate in an experiment. Beforehand the parents of the children had signed consent forms. We told the children that we wanted to know how they would search using this website, and gave them their individualized tasks. We started the screen recording, after which the children could start. They were told that they could ask us questions any time they wanted. When they finished answering their tasks, the questionnaire was given and the next child was retrieved.

3.2 Results

The children that participated use the computer an average of a couple of times a week. There was no difference between the grades. Most children use Google (70%) or Wikipedia (43%) to search (the children could give multiple answers to this question). Based on the collected data and our experience during the experiment, we anticipated that there would be a difference between age groups in the results. Therefore, we also looked at differences between the two grades. We divided our main results in four parts: the time the children used, how they navigated, what kind of results they chose, and based on the search behavior of the children we defined five types of searchers.

3.2.1 Time

Table 3 shows the time it took the children to complete the tasks. To be more precise we looked at total time to complete the task but also at the duration of the navigation, and the result selection time: how long it took the children to come to an answer after finding the correct topic.

	Mean 7th grade	**SD**	**Mean 8th grade**	**SD**	**t(df)**	**Signific. t-test**
Task 1						
Total time	545	356	307	183	t(34.5)=2.76	*p*=0.009
Navigation time	113	109	43	37	t(28.4)=2.87	*p*=0.008
Result selection time	432	341	263	180	-	ns
Task 2						
Total time	347	201	275	174	-	ns
Navigation time	123	103	158	161	-	ns
Result selection time	224	178	117	106	t(37)=2.19	*p*=0.035
Task 3						
Total time	1046	356	716	371	t(38)=2.85	*p*=0.007
Navigation time	139	215	122	166	-	ns
Result selection time	908	304	595	269	t(38)=3.38	*p*=0.002

Table 3: Mean time (in seconds) to complete the tasks and the results of an independent t-test to test significance of differences between grade 7 and grade 8.

For all measurements we calculated if there was a significant difference between the two grades ($p < 0.05$, see Table 5).The children of grade 7 did take significantly more time to complete task one and task three. When looking more precisely at the navigation time we see that 7th grade children only took significantly longer in the first task, while the time they needed to find the final result they used for answering was longer in the second and third task. For the first task this difference was nearly significant (p=0.051).

3.2.2 Navigation

We define navigation as the path that was taken to the correct topic for answering the task. The place children start is also counted, because this can already be the correct place to look. Also the selection of the correct topic is seen as a step. We annotated the number of navigation steps.

Afterwards we looked at looping of the same steps. We did this for all tasks. The results are presented in Table 4.

	Number of navigation steps		Percentage of looping	
	Mean (SD) grade 7	**Mean (SD) grade 8**	**Grade 7**	**Grade 8**
Task 1	4.9 (4.9)	3.1 (2.3)	26%	18%
Task 2	5.7 (5.1)	4.2 (2.8)	36%	18%
Task 3	6.3 (6.8)	6.8 (6.8)	22%	47%

Table 4: Number of navigation steps and occurrence of looping behaviour in the navigation separately for grade 7 and grade 8.

For the first task the most efficient route consisted of 2 steps, for the second task this was 3 steps, and for the last task it could be 2 or 3 steps depending on the topic of the task.

In all tasks the mean number of navigation steps was higher for the 7th grade children than for the 8th grade children. When looking at looping behaviour we see that the 7th grade children displayed more looping behaviour in the first and second task and 8th grade children displayed more looping behaviour in the third task. For all measurements we used the independent t-test to calculated if there was a significant difference between the two grades ($p < 0.05$). There were no significant differences between grade 7 and grade 8 in the number of navigation steps and the percentages of looping in the navigation

3.2.3 Selection of results

For all tasks we annotated the number of results that were selected after finding the correct topic for answering the task. Afterwards we looked more in-depth at the results. We identified looping behaviour in the selected results; which means that children looked multiple times at the same results. We investigated how many times the first result was selected first, and how many times a Wikipedia result was chosen. The choosing of the Wikipedia results is interesting, because these results had no image but the text “geen plaatje” (no image) instead. The results are presented in Table 5

	Nr. of selected results		Looping in results		First result sel. first		Wikipedia selection	
	Mean (SD) grade 7	**Mean (SD) grade 8**	**Grade 7**	**Grade 8**	**Grade 7**	**Grade 8**	**Grade 7**	**Grade 8**
Task 1	4.8 (4.0)	3.2 (2.2)	30%	24%	57%	53%	74%	41%
Task 2	3.4 (3.1)	1.9 (1.7)	23%	12%	77%	24%	-	-
Task 3	4.0 (2.7)	3.5 (2.5)	35%	35%	74%	35%	65%	71%

Table 5: Number of selected results, looping in results selection, first result selection and Wikipedia selection separately for grade 7 and grade 8.

For all tasks, the average number of selected results was higher for the 7th grade children than for the 8th grade children. However, according to the independent t-test using a significance level of 0.05 again, the differences were not significant. In task 2 and task 3 grade 7 children chose the first result significantly more often than grade 8 children (task 2: chi-square=11.15, *df*=1, *p*=0.001; task 3: chi-square=5.96, *df*=1, *p*=0.024). This might indicate that the older children looked at the result list of the internet search more critically. There were no significant differences in looping behaviour and selection of Wikipedia results either. Notice that the Wikipedia results were selected very often by children of both grades. In the first task, children of grade 7 chose more Wikipedia results than children of grade 8 but this was only marginally significant. (chi-square=4.37, *df*=1, *p*=0.053).

We also looked more specifically at the selected results to see if there was a difference between the two grades concerning the type of selected results. We looked at results with a correct image, results with an incorrect image (an image not matching the topic that was searched for) and results with a missing image and the text “geen plaatje” (no image) instead. Task-answers that

were retrieved making use of a backdoor, or without selecting a result, were put in the category remaining. An example of a backdoor is a child that uses the search bar in Wikipedia to search for the answer of the task. This can lead to finding the correct answer, however without using the website. It was possible to give an answer to some tasks without selecting an image because of the general information displayed about the topic when selecting it in the topic-bar. This possibility also fell in the category remaining.

The percentages for the four categories were calculated by looking at the total number of results that all children selected in a specific category(correct image, without image, incorrect image, and remaining). This number was divided by the total number of results that all children selected during the specific task. The results are presented in Table 6.

	Percentage of chosen images	
	7th grade	**8th grade**
Task 1		
Correct image	45%	25%
Without image	23%	16%
Incorrect image	2%	15%
Remaining	31%	44%
Task 2		
Correct image	52%	45%
Without image	9%	0%
Incorrect image	7%	0%
Remaining	32%	55%
Task 3		
Correct image	24%	29%
Without image	24%	25%
Incorrect image	21%	12%
Remaining	31%	34%

Table 6: Percentage of chosen images in the four categories, for each task separately and the results of an independent t-test to test significance of differences between grade 7 and grade 8.

We calculated whether there was a significant difference between the two grades in type of results they selected. To make a good comparison we calculated the percentage of each category for each child. For example, if a child chose five images in total, of which two images were correct, then this was coded as 0.4, indicating that 40% of the chosen images were correct. We did this to take into account that some children chose few images, while others chose more.

Looking at the differences between the grades in selected images we see that in the second task 7th grade children selected more results with a correct image: grade 7: 44%, grade 8: 24%. This difference is significant ($t(26.1)=2.32$, $p=0.028$). Although not significant we see the same trend in the first task, while in the third task the difference is almost non-existent. 7th Grade children also selected significantly more results with a missing image in the second task (grade 7: 6%, grade 8: 0%, $t(21)=2.46$, $p=0.023$). Although missing images were mostly related to Wikipedia results, the second task did not have a Wikipedia result. More 8th grade children answered the second task by using the general information displayed about the topic (grade 7: 47%, grade 8: 76%,

t(30.4)=-3.10, p=0.004). There was no significant difference between the two grades in selected results with an incorrect image.

3.2.4 Searcher types

We wanted to further analyse the different search behaviours the children displayed. Therefore we defined five searcher types. The properties of the searcher types are loosely based on Druin et al. (2010). We can not use their so-called search roles because they base them on keyword interfaces, and their setting was completely different. We combined the navigation, the selection and screen recordings to come to our five searchers types. The screen recordings were seen as more important than the number of navigation steps and the selection, because the recordings gave the context of the navigation steps and the result selection.

Our five searcher types are the following: a chaotic searcher, an unstructured searcher, a distracted searcher, an explorative searcher, and a directed searcher. We will first give a short explanation with some characteristics of each type.

Chaotic searcher. As the name says, this searcher's most important characteristic is his chaotic search behaviour. He is quick to try another result and quick to decide that the answer cannot be found in the section he is looking at. He begins by trying the first thing, but before he has tried it, already switched to the next thing. The goal of his search might not be completely clear to him.

Unstructured searcher. This searcher is similar to the chaotic searcher, with the difference that he switches less quickly to try something else. This searcher has a clear goal in mind but is not structured in fulfilling it. This searcher is more likely to look in illogical places.

Distracted searcher. As the name implies, this searcher's most important characteristic is that he is easily distracted. This happens when he sees another image that he finds interesting. He will easily start to further investigate this subject.

Explorative searcher. This searcher's most important characteristic is that he does a lot of browsing. He explores the website and its structure, in order to find what he is looking for.

Directed searcher. This searcher has a clear goal in mind, is effective and tries to avoid unnecessary steps. He is not distracted by other interesting-looking images and does less browsing than the explorative searcher. He tries to think where to look before acting.

We mainly used the screen recordings to determine the type of searcher of each child. These numbers can be found in Table 7. The percentage is calculated by dividing the number of chaotic searchers in a grade by the total number of children in the grade. Note that a child can change its search behaviour and can therefore be different types in one session. Therefore, the percentage calculated is the percentage of children that display the behaviour associated with the type within a grade. So 94% of the 8th grade children display behaviour that is associated with the directed searcher in all the three tasks.

	Task 1		**Task 2**		**Task 3**	
	7th grade	**8th grade**	**7th grade**	**8th grade**	**7th grade**	**8th grade**
Chaotic searcher	22%	12%	14%	0%	22%	0%
Unstructured searcher	13%	0%	14%	12%	17%	12%
Distracted searcher	17%	18%	5%	12%	4%	0%
Explorative searcher	43%	6%	9%	6%	26%	29%
Directed searcher	61%	94%	77%	94%	65%	94%

Table 7: Percentages of annotated searcher types. A child can change its search behaviour during a session hence one child can belong to different searcher types.

We used Fisher's exact test to calculate whether there was a significant difference between the two grades for each searcher type. During the first task, the 7th grade children displayed significantly more explorative search behaviour than the 8th grade children (p=0.012). Search behaviour that is related to the directed searcher is mostly displayed by children in the 8th grade. The difference is significant and most strong in the first task (p=0.026). In the third task we see the same difference, although only marginally significant (p=0.054). Children of both grades get distracted sometimes, but that behaviour does not occur very often and there is no significant difference between the grades. In the first task, 7th grade children are more unstructured, but this is not significant. Although only marginally significant, during the third task the 7th grade children were more chaotic in search behavior (p=0.061) than the 8th grade children who did not show chaotic behavior at all.

3.3 Discussion

When looking at the results we see that there is a difference between the 7th and 8th grade students. Before doing the experiment we did not expect this difference to be this notable. However, during and after conducting the experiment differences became apparent. Therefore, we analysed these differences in the data. The differences between the two grades are visible throughout the whole results sections. For example, the 8th grade students took significantly less time to complete the first and third task. Other results also point in the direction that there is a larger difference in development than we first expected. According to Piaget children of age 7-11 learn to think logically about objects and events, they achieve a notion of the conservation of number, mass and weight, and can classify objects according to several features. From 11 years and on they learn to think logically about abstract concepts (Cooper, 2005). Most of our participants were 11 years of age. We think that the grade the children are in is decisive whether children are able to search more effectively, especially when they are about the same age. It is clear that in this age range one year (or grade) can be of huge difference in performance.

The navigational route the children took is not the most efficient route. However, when keeping in mind that this was an unknown website for them, and they probably did not search image based before it is not bad. Two children from grade 8 did even take the most efficient route in all tasks.

Looping behaviour occurred in the navigation as well as in the selection of the results. There was no significant difference between the two grades in this behavior. Looping may occur when a user does not recall the hyperlinks visited or the searches executed, or when a user decides to revisit previously retrieved results. Recall causes a cognitive load for all types of tasks in information retrieval systems (Borgman et al., 1995), while children have limited recall knowledge (Siegler, 1991). Looping behaviour is found in other studies with fact-based tasks as well (Bilal, 2000), however it is not clear to which extend the type of tasks is also related to the looping behaviour. We think that looping behaviour might depend on the task. This task determines whether looping behaviour might be beneficial, for example: in an open-ended task looping might also indicate that a child is trying to combine the information of multiple sources. However, we do not have any evidence for this theory.

We also looked more specifically at the selected results. Results with a correct image were chosen often, even if the result itself was not that good. An example of this was the fourth result in the second task. This was an excellent picture of a Neanderthal; however the result itself was not good. Still it was chosen, and this result was more than the other results subject to looping behaviour. Most of the missing images were Wikipedia results, which could be seen in the label. We think that this is the main reasons why the missing image results were chosen. Apart from the second task there is not a significant difference in the selection of results with a missing image between the two grades. This is interesting to see because in the second task the result with the missing image was not a Wikipedia result. However, it is also the task in which the children of both grades did select less results with a missing image. We do not have a satisfying explanation for this, although it might be that the 7th grade children expected this result to be a Wikipedia result and did not read well.

The first result was significantly chosen more often in the second and third task by 7th grade children than by 8th grade children. It looks like the 7th grade children expect the first result to be the best, while the 8th grade children give less value to this.

After looking at navigation and selected results we tried to get more insight in what kind of searchers children are. For this we defined five searcher types. We combined the navigation, the selection and screen recordings to come to these five searchers types. The screen recordings were seen as more important than the number of navigation steps and the selection, because the recordings gave the context of the navigation steps and the result selection. When looking at the searcher types the first thing we see is that again there is a difference between the two grades. During the first task 7th grade children tend to explore the structure of the website more, and thus display more explorative behaviour. The difference is obvious in the directed searcher type, which is displayed by 94% of the 8th grade children during all tasks, which is a very high precentage. We can also see that the 8th grade children display less chaotic or unstructured behaviour. Hence, with this image-based search interface, children of grade 8 are quite effective searchers with a clear goal in mind.

3.4 Conclusion

In this study we investigated how children search with an image-based interface. We looked at the kind of searchers children are when using this interface. Do they display chaotic search behaviour and become distracted by all the interesting images? Or do they have a clear goal and use the images effectively to find the answer? What is important when designing an image-based search interface? Is the quality of the image of importance?

The quality and accuracy of the images is important in an image based interface. Results with a correct image were chosen more often, and were more likely to be a subject in looping behaviour. While the last is something to avoid in some tasks, it is clear that the quality and accuracy are of influence.

We found a larger difference in behaviour and performance between the two grades than we anticipated beforehand, which was especially clear in the time it took to complete the tasks. With this image-based search interface older children (age 11-12) needed less time, displayed less chaotic and unstructured behaviour and were directed searchers: effective and with a clear goal in mind. We conclude that in this age range one year (or grade) can make a big difference in search behaviour and performance.

4 Evaluation of the Hospital Demonstrator – Stage 1

To improve the accessibility of medical information for children, we developed the PuppyIR Hospital Demonstrator for the Emma Kinderziekenhuis (EKZ) at the Amsterdam Medical Centre (http://www.emmakids.nl/). The Hospital Demonstrator, also known as the Emma Search (EmSe) engine, has the goal to improve the accessibility and services of the Patient Information Centre (PIC) by: (1) providing an engaging interface that encourages children to explore and learn, (2) facilitating query formulation, (3) improving the understandability of content, and (4) enabling moderated and trusted web and medical site search services. An overview of the Hospital Demonstrator and its features is provided in Deliverable 7.4 (Hospital Demonstrator – version 2.0).

The first version of EmSe was released in early 2012. It is envisioned to be accessible to staff and patients within the hospital (via bedside and other terminals in the hospital), and also to out-patients via the web. The evaluation consisted of three main stages. (1) The first stage was to obtain feedback from the staff at the hospital. From this initial evaluation we refined the demonstrator to incorporate suggested changes before the first patient contact. (2) In the next stage, we focused on the patients. We actively solicited feedback from patients This stage followed the set-up described by Dekker (2011), in which the author conducted a user study, asking elementary school children to perform various search tasks. The EKZ's in-house school is an optimal venue for the replication of this study in order to compare the usefulness of EmSe with that of popular general Web search engines such as Google or Bing. (3) Finally, in a more long-term perspective, we will use built-in interaction logging mechanisms to get a better perspective of how and how frequently the system is used. In this section we present the results of stage 1 of the evaluation. The results of stage 2 will be presented in section 5.

In March 2012, we visited the children's hospital and interviewed a total of 11 staff members in 8 individual sessions. Table 1 details the professional background of the interviewed participants. At the start of each session, the participants were shown a brief explanatory video, introducing the features and functionality of EmSe. Afterwards they were encouraged to use the demo themselves. Due to scheduling reasons, some participants interacted with the interface in groups rather than individually. The participants were encouraged to think aloud while operating the system and the interviewer did not interfere with their search activities unless directly asked for assistance or clarification. Afterwards, all participants were asked for their general opinion on the system and any aspects that they particularly liked, disliked, or missed. Finally, they were asked to comment on major aspects such as Search result quality, interface design, the Body Browser and the content simplifications. Sessions typically took 15-20 minutes each.

Professional background	Frequency
Care & nursing staff	3
In-house teachers	2
Pedagogues	3
Pediatricians	2
PIC staff	1

Table 1: Participant demographics of the EmSe evaluation, stage 1.

There were a number of minor bugs as well as localization issues (the interface is Dutch and not all developers were Dutch native speakers) that were spotted in this stage of evaluation. They have been addressed in the current version of EmSe. In the following, we will discuss the participants' feedback concerning EmSe's main components as well as additional features that were suggested and that remain to be investigated in the future.

4.1 Search result quality

The participants all judged the search result quality high. Some suggested to more explicitly, e.g., in the form of result re-ranking, separate general web results for adults from those that were explicitly designed for children. Additionally, there was a concern that some topics such as death or cancer should only be reachable when explicitly being queried for. This form of search moderation represents one of the major challenges for future additions as it involves a) identifying sensitive topics that require special treatment, and, b) based on query, user context and session information, decide whether or not to display sensitive information.

4.2 Interface design

While participants liked the clean uncluttered interface, they encouraged further personalization and integration of the user into the search process by adopting a login concept so that the user could be individually addressed by the system. This could for example happen in the form of the puppy avatar (representing the search engine) calling the child by its name in the search prompt and the interactive steps such as query modification.

An additional request for mobile device compatibility of EmSe was already partially addressed in the latest version which is fully functional on systems like the IPad. Full functionality on small-screen mobile devices is currently not supported.

4.3 The Body Browser

The body browser was unanimously liked by all interviewed staff members. They considered it one of the key features as it would facilitate browsing-driven information discovery. They suggested a number of additional anatomical structures (e.g., the appendix or the genital tract) to be included for greater coverage.

As the Body Browser opens as an overlay to EmSe's default screen, several users were confused to not be able to jump back and forth between Body Browser and search results using the browser's back button. Future versions will consider enabling this browsing mode in order to enable a more natural interaction pattern.

4.4 Content simplifications

The participants found in-line content simplifications helpful and non-intrusive. It was suggested to offer the simplification service independently of a term's expected diffiulty in order to account for different user preferences and needs. More concretely this would mean to not highlight difficult terms but rather offer an interface through which arbitrary term simplifications could be requested on-line.

4.5 Conclusion

The participants all judged the search result quality high, they liked the clean uncluttered interface and they found the in-line content simplifications helpful and non-intrusive. All interviewed staff members unanimously liked the body browser they considered it one of the key features as it would facilitate browsing-driven information discovery.

They also had some suggestions for additions. They encouraged further personalization and integration of the user into the search process, they would like to be able to jump back and forth between Body Browser and search results in order to enable a more natural interaction pattern, and they requested mobile device compatibility of EmSe. Finally there was a concern that some potentially frightening topics should only be reachable when explicitly being queried for.

The minor bugs that were spotted in this stage of evaluation have been addressed in the current version of EmSe that has been evaluated with children (see section 5). Feedback concerning

EmSe's main components as well as additional features that were suggested remain to be investigated in the future.

5 Evaluation of the Hospital Demonstrator – Stage 2

The second stage of the EmSe evaluation was conducted on May 15th, 2012, involving patients of the EKZ. 6 patients of ages 10-17 were presented with EmSe (http://wickham.dcs.gla.ac.uk:8080/hospital/) and Microsoft Bing (http://bing.nl) in order to solve a number of medical information finding tasks.

5.1 Setup of the experiment

All participants were Dutch native speakers and the interaction with the researchers as well as the search efforts were conducted in Dutch. At first, the participants received a brief introduction into the EmSe search system and its functionalities such as the Body Browser, suitable query suggestions or content simplifications. Subsequently, we compared the usefulness of EmSe with Microsoft Bing. Starting on one of the systems (the order was alternated, half of the participants started with Bing, the other half with EmSe) they were given first a closed-class question (simple factual "What is x?" type question), then an open-ended one ("List all information that you can find about x."). The same procedure was repeated on the second system (with different questions). Finally, each participant was asked to search for information concerning their own medical condition using their preferred choice among the two compared systems. Using a set of fall-back questions from both types, we made sure that no participant encountered questions concerning their own medical conditions prior to this final question. To conclude each session, the participants were asked to fill a brief form in which they could describe their search experience. The questionnaire in this form is part of the user evaluation toolkit on the PuppyIR site (see http://www.puppyir.eu/results/user-evaluation-toolkit). Individual sessions lasted for approximately 40 minutes. The researchers took notes concerning search behaviour and success during the full duration of the sessions and provided support where necessary, otherwise avoiding interference with the participants' search efforts.

The questions offered belonged to the following set:

Closed questions:
1) Wat is een Port-a-Carth? (What is a port-a-carth?)
2) Wat id een Mic key button? (What is a mic key button?)
3) Wat is een PEG? (What is a PEG?)

Open questions:
4) Wat kun je vinden over de taaislijmziekte?
(What information can you find about cystic fibrosis?)
5) Wat kun je vinden over de ziekte van Crohn?
(What information can you find about Crohn's disease?)
6) Wat kun je vinden over de sikkelcelziekte? (What information can you find about the Sickle-cell disease?)

5.2 Results

During the six sessions of the study, participants were generally positive about EmSe's look and feel as well as the quality of the presented results. The Body Browser was unanimously liked and inline content simplifications were considered useful. 100% of participants chose EmSe to answer the final question about their own medical condition, irrespective of the order in which the two search engines were presented to them.

Due to network problems, the server on which EmSe was offered, was not always responsive. Two participants commented on the system being slow. Query suggestions were frequently (50% of the sessions) confused with search results. The participants afterwards stated that they would have expected suggestions to be offered in the form of automatic completions as done by several

commercial search engines. Older participants (aged 13+) often found the interface too focused on young children. Since the target audience of EmSe are children of ages 8-12, this was to be expected. Especially for children with less developed reading and writing skills, the content simplifications were helpful. Interestingly enough, this feature was frequently used by older participants as well.

The concluding questionnaire supported a number of previous assumptions, such as older children being more likely to be attracted by conventional search engines such as Bing. At the same time, a number of surprising observations could be made: (1) Older participants liked the inline content simplifications more than their younger peers. (2) Bing was reported to be easier to use than EmSe. We attribute this to the participants already being used to the search interfaces and paradigms of conventional web search engines. Every participant reported using Google for their day-to-day information needs. (3) Despite their familiarity with Bing's interface, EmSe consistently received higher agreement scores for the statements "I like this system" and "I would like to use this system again". These findings should be treated as trends and indications rather than hard facts, as statistics over a population of 6 participants cannot be considered very robust.

5.3 Conclusion

In the first 2 evaluation stages, we gained a qualitative understanding of professionals' and patients' usage behaviour and wishes for modifications towards EmSe. In both stages, there was a strong consensus about the system's usefulness for children's information search. A number of shortcomings were discovered and either directly addressed or discussed for later integration into subsequent versions of EmSe. Most of these requests concerned the interface as users expected interaction paradigms they are used to from commercial web search engines. For productive use of EmSe a closer orientation towards well-known behaviour such as displaying query suggestions in the form of automatic completions might be advisable. Greater numbers of participants are aimed to be reached in the analysis of user interaction logs (evaluation stage 3). They are needed to strengthen the conclusions drawn from the small-scale evaluations of stages 1 and 2.

6 Conclusion

The results of the evaluation of the Museon part of the Museum Demonstrator did not show the expected positive influence of the multi-touch table application on collaboration and enjoyment. However, the results on the enjoyment scales were quite high. both for groups that started the museum visit with the multi-touch table and for groups that only used the table in the end.

We found a larger difference in perceived enjoyment, behaviour and performance between the two grades than we anticipated beforehand: children of grade 7 often scored higher than children of grade 8 on enjoyment as measured by the Smileyometer and the Children IMI interest/enjoyment scale. When working with the image-based website at the school, older children (age 11-12) needed less time, displayed less chaotic and unstructured behaviour and were directed searchers: effective and with a clear goal in mind. We conclude that in this age range one year (or grade) can make a big difference in perceived enjoyment, search behaviour and performance.

An important result of this evaluation of the Museum Demonstrator is the fact that we adapted the IMI scale for adults to get a reliable scale to measure intrinsic motivation of children of 10-12 years old.

In the evaluation of the Hospital Demonstrator the professional participants (hospital staff) all judged the search result quality high, they liked the clean uncluttered interface and they found the in-line content simplifications helpful and non-intrusive. All interviewed staff members unanimously liked the body browser. They considered it one of the key features as it would facilitate browsing-driven information discovery.

In the Hospital Demonstrator evaluations with patients, the children were generally positive about EmSe's look and feel as well as the quality of the presented results. The Body Browser was unanimously liked and inline content simplifications were considered useful. All participants preferred EmSe to answer the final question about their own medical condition. However, most users expected interaction paradigms they are used to from commercial web search engines. For productive use of EmSe a closer orientation towards well-known behaviour such as displaying query suggestions in the form of automatic completions might be advisable.

References

Bilal, D. (2000). Childrens use of the yahooligans! web search engine: I. cognitive, physical, and affective behaviors on fact-based search tasks. Journal of the American Society for Information Science, 51(7), pp. 646-665.

Borgman, C. L., Hirsh, S. G., Walter, V. A. and Gallagher, A. L. (1995). Childrens searching behavior on browsing and keyword online catalogs: The science library catalog project. Journal of the American Society for Information Science, 46(9), pp. 663-684.

Cooper, L. Z. (2005). Developmentally appropriate digital environments for young children. Library Trends, 54(2), pp. 286-302.

Dekker, P. (2011). Children's roles in web search. Master Thesis Delft University of Technology.

Druin, A., Foss, E., Hutchinson, H., Golub, E. and Hatley, L. (2010). Children's roles using keyword search interfaces at home. In Proceedings of the 28th international conference on Human factors in computing systems, CHI '10, pp. 413-422, ACM, New York, NY, USA.

Read, J.C. and MacFarlane, S. (2002). Endurability, engagement and expectations: Measuring children's fun. In Proceedings of International Conference Interaction Design and Children (IDC 2002), pp 1–23.

Siegler, R. (1991). Children's Thinking. Prentice Hall, Englewood Cliffs, New Jersey.

University of Rochester. Intrinsic motivation inventory. Retrieved May 10, 2012, from http://www.selfdeterminationtheory.org/questionnaires/10-questionnaires/50

Yiannoutsou, N., Papadimitriou, I., Komis, V. and Avouris, N. (2009). Playing with Museum Exhibits: Designing Educational Games Mediated by Mobile Technology. In IDC '09: Proceedings of the 8th International Conference on Interaction Design and Children, pp. 6–9.