Evaluating serious games with young children

Evaluation methods and research

Maria Skaaden
Department of Product Design
Norwegian University of Science and Technology

ABSTRACT

Many serious games are developed specifically for young children, and must be tested with children. Testing and evaluating serious games ensures their quality. The purpose of the research reported in this paper is twofold, to present an overview of existing literature with regard to evaluation methods, and to assist the evaluator when choosing and designing an evaluation method.

The paper is based on a literature search focusing on evaluation methods for young children and serious games. Evaluation methods for serious games for children should take into account children’s needs and range of cognitive abilities. Serious games have specific purposes besides being entertaining. Their purpose can be to promote learning, teach skills and behaviour, and even reduce feelings of stress and anxiety. Evaluation of such games should therefore consider the effectiveness of the game. The paper gives an overview of existing evaluation methods for usability testing with young children. A brief introduction of serious games is given to describe aspects of serious games that can be evaluated. Characteristics of evaluation methods, characteristics of young children and aspects of serious games are summarized in a table. The findings can inform the choice and design of an evaluation method.

KEYWORDS: Evaluating, methods, serious games, young children

1. INTRODUCTION

Today’s children are growing up in a digital world. Children as young as 12 months are able to use the iPad, and children are choosing to play digital games for hours during their leisure time. The App-store is full of games for young children (Shuler, 2009, 2012). Young children are a special user group. They are especially vulnerable to media messages (Lieberman et al., 2009) and their needs and interests, as well as developmental abilities should be addressed in the design and evaluation of digital games.

Serious games are finding their way into homes as well as institutions like schools and hospitals (Lieberman, 2001, 2009). Serious games include games for learning, interactive health games and games for skill and behaviour learning. Serious games have a specific purpose besides being merely fun or entertaining, which make them especially important to evaluate. If they do not fulfil their purpose, they are not effective.

In the paper “Examining Values: An Analysis of Nine Years of IDC Research” a content analysis of
the values expressed within the research community of interaction design and children (IDC) is done. It is found that the majority of IDC papers were targeted to children between 6–12 years, there were no papers focusing on children under the age of two. Young children have different needs and may require new ways of designing and evaluating technologies. The paper concludes that: “IDC can design for a larger variety of ages, rather than focusing primarily on children ages 6 – 12. They could be more explicit about their theoretical perspectives and use more theories and models from other fields.” (Yarosh et al., 2011)

Findings from literature search show that few researchers investigate usability-testing methods with children under the age of 5. While many articles focus on children and digital games and computer games, few researchers write about evaluation of serious game applications for young children. There is a need for more insight into evaluation methods for serious games for young children, so the games that are designed can better fit children’s needs and fulfil their purpose.

The purpose of the research reported in this paper is twofold, to present an overview of existing literature with regard to evaluation methods, and to assist the evaluator when choosing an evaluation method. Choosing and designing an evaluation method is also a part of the overall design process, and it is found useful to give an account of other researchers’ approaches.

1.1 Structure of the paper

The first part of the paper gives an introduction to serious games.

The following section focuses on what the evaluator should have in mind when carrying out usability testing with young children, and provides a short overview of young children’s development, their cognitive abilities and needs.

Children’s role in the design process and the need to include them is also briefly discussed.

The paper goes on to examine different usability testing methods for young children, and research on evaluation methods. Finally, the findings are summarized and discussed.

1.2 Research methodology

Research and methods were found by literature search.

<table>
<thead>
<tr>
<th>Research on evaluation methods</th>
<th>Research on child development</th>
<th>Research on serious games</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inform the choice and design of an evaluation method</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Fig. 1: Research approach*

Research on serious games help identify aspects of serious games that can be evaluated. Research on child development gives insight to characteristics of children to consider when doing usability testing. Finally, research on evaluation methods gives an overview of existing methods, and characteristics of evaluation methods. Research regarding assessment of evaluation methods is also cited.

2. EVALUATING SERIOUS GAMES

2.1 Children and digital media

Carly Shuler researches digital media and children’s learning. In a recent report (published January 2012) Shuler presents an analysis of the educational content of the App Store. Shuler found that “Over 80% of the top selling paid apps in the Education category of the iTunes Store target children.”

Other important findings:
- The percentage of apps for children has risen in every age category, accompanied by a decrease in apps for adults.
- Apps for toddlers/pre-schoolers are the most popular age category (58%), and experienced the greatest growth (23%). General early learning is the most popular subject category.
Educational apps clearly have a big market share of the App Store. The designers, developers and evaluators have a responsibility to ensure the quality and “child-safety” of these apps. Evaluating the design with experts on children or children throughout the design process will do just that.

Children’s needs, abilities and relation to technology change while growing up. “A successful study of their interaction with technology requires that we understand the purpose and context for which children will interact with technology as well as their own needs” (Markopoulos and Bekker, 2003a).

2.2 What is a serious game?

Lieberman et al defines digital games as “rule-based systems that involve challenge to reach a goal while providing feedback on the player’s progress. The rules, challenges, and feedback of digital games can create deeply engaging experiences” (Lieberman et al., 2009). In “Serious Games - An Overview” (2007) Susi, Johannesson and Backlund discuss the concept of serious games by looking at different definitions and descriptions of the concept. They found that there are many different terms that point to the concept of serious games. Although there are many definitions, “One issue most definitions agree upon, more or less, is that serious games are concerned with the use of games and gaming technology for purposes other than mere entertainment or “fun”. Such purposes include education, training, health, etc. Although fun and entertainment is excluded in many authors’ definitions, or used for drawing a line between serious and other games, others argue they constitute key components of serious games.” (Susi et al., 2007).

Serious games are some place in the middle between entertainment games and software tools. Software tools are used for different purposes, e.g. Word is used for writing text. When evaluating software tools, usability is the most important aspect. Serious games have specific purposes (e.g. learning) that other games have not, but that does not imply that serious games should not be fun. What would a game be without fun? Who would want to play a game that is not fun? The quality of any digital game depends both on usability and fun. Pagulayan et al (2003) wrote, “The ease of use of a game’s controls and interface is closely related to fun ratings for that game. Think of this factor as the gatekeeper on the fun of a game”.

2.3 Different categories of serious games

Some categories of serious games for children include:

**Health games**
- Games for training
- Disease management (asthma, cancer, diabetes)
- Distraction from pain and anxiety during medical treatments

**Education and learning**
- Educational games in schools
- Cognitive skills (memory, visual attention etc.)

**Social interaction**
- Cooperation

Sources: (Lieberman et al., 2009), (Lieberman, 2001), (Gerling et al., 2011)

2.4 Why is there a need for methods to evaluate serious games?

Serious games have specific purposes, besides fun and entertainment. The purpose of the game varies from different categories of serious games, but should always be the core of the game. The purpose could be learning, teaching skills and behaviour or even pain distraction. “Poorly designed games can be time-wasting sedentary activities that contribute little to children’s...
learning, skill building or healthy development” (Lieberman et al., 2009). Very poorly designed games can do harm by teaching unwanted lessons, create aggressive or anti-social behaviour, and generate fear or anxiety. If serious games are not well designed, and do not have their purpose at the heart, they could be harmful for young children. Testing and evaluating games is a way to ensure their quality. Usability testing with children will provide important insights to children’s opinions of the game, whether or not the game meet children’s needs and cognitive abilities, as well as the effectiveness of the game.

2.5 What can be tested?

When choosing and designing a method for usability testing serious games it is important to consider which aspects of the game that should be tested.

Aspects of the game that could be tested:
- Usability
- Fun
- Use in real context
- Use over time
- Effectiveness of the game (e.g. learning)
- Likeability, children’s opinions
- Appeal over time
- Emotional states
- Age-appropriateness
- Challenge, curiosity, control, fantasy (Malone and Lepper, 1987)

3. USABILITY TESTING WITH YOUNG CHILDREN

Conducting traditional usability testing with young children can be challenging. Young children cannot conduct a task for very long, and are easily distracted. They have difficulty expressing their likes and dislikes verbally, and try to please adults. They could tell you they like the product being evaluated just to please you (Hanna et al, 1997). Table 1 in the appendix presents cognitive abilities of young children at different ages, and can be useful for the evaluator.

When designing and evaluating products for children it is often necessary to make some assumptions about a group of children with respect to their abilities and skills at a given point in time (Markopoulos et al., 2008). Learning about children, their development, abilities and needs can provide the evaluator with useful insights.

3.1 Learning about children

The easiest way to learn about children is to spend time with them. If you do not have children on your own, spending time with children of friends and family is a possibility. Watching children’s TV-programs, reading children’s books, and observing what things they like to play can be revealing. Educational curricula in the country, which you are working, can be a good indication of children’s abilities at different stages (Markopoulos et al., 2008).

3.2 Children’s development

Child development is the study of the differences and similarities of children across a similar population. Piaget’s theory of development (1970) has had a significant impact on education and developmental psychology, and is still widely used. Although the theory has been criticized it is still a good starting point when researching children’s development.

3.3 Piaget’s stages of development

3.3.1 Sensorimotor: birth – 2

Children experience the world through movement and senses. During the sensorimotor stage children are egocentric, meaning they cannot perceive the world from others’ viewpoints. Language and thought processes are limited (Guinard, 2000).
3.3.2 Preconceptual thought: 2 - 4

During years 2-4 children cannot yet manipulate and transform information in logical ways, but they can think in images and symbols. The child is able to name objects that are not present, but thought of. Other examples of mental abilities are language and pretend play.

3.3.3 Intuitive thought: 4 - 7

Children tend to grow very curious and ask many questions; they begin the use of primitive reasoning. There is an emergence in the interest of reasoning and wanting to know why things are the way they are. Piaget called it the Intuitive thought-stage because children realise they have a vast amount of knowledge but do not know how they know it. Children can use symbols and words and can distinguish reality from fantasy. In the latter part, they can take into account the viewpoint of others (Markopoulos et al., 2008).

There is a lot of variation within age groups. A 5-year old can behave as a 7-year old, and vice versa. Development is a continuous process.

3.4 Cognitive developmental

Cognition includes intellect and language. The way individuals learn to think, their memory, their problem-solving abilities, and their reasoning. The focus in the cognitive-developmental perspective is on the child’s mind. Piaget states that children act as scientists, and in doing so actively construct meaning as they discover how the world works.

3.5 Sociocultural

Bronfenbrenner’s theory (Ecological Systems theory) is the most frequently used contextual view of development. His view is one of a child developing within a complex system with many interactions (Markopoulos et al., 2008).

3.6 Socioemotional

The most important aspect of socioemotional development is an appreciation of how relationships affect development. The self-esteem of children is associated with emotional development. Preschool children have very high levels of self-esteem, and with time (possibly as a result of entry into the educational system) this self-esteem declines (Markopoulos et al., 2008).

3.7 Children’s role in the design process

While user-centred design includes children as users and testers, participatory design gives children the role as design partners. Druin is an advocate of participatory design and including children at all stages of the design process. Druin has developed a method called cooperative inquiry where children act as design partners, which is the highest level of involvement. The figure shows the levels of involvement and roles children can have in the design process.

Fig. 3: The relation of children to technology can vary simply being the end-user of a technology that is designed by adults, to a very active involvement as a member of the design team. Figure adapted from Druin (2002).

The design process is an iterative process. In the first phase of the design process designers rely on creative methods to gain insight into
children’s needs and interests, likes and dislikes. Examples of creative methods are cooperative inquiry, drawings, mood boards and low-tech prototyping. Further on in the design process some test material is usually available; sketches, mock-ups or prototypes.

In these later stages evaluation and benchmarking of the design becomes more important. The designer needs to assess if user requirements have been achieved and detect usability and fun problems. The role as design partner changes during the design process, just like a designer’s focus or aim changes during the design process. Children can also be involved only at the later stages, then as users and testers.

3.8 Why should children be involved?

Adults and children view the world differently, and adults may not know what children like and want. Children have their own environments (Druin, 1999), which adult researchers need to understand when designing and evaluating children’s products. In the process of designing technologies for children there is a range of participation for children from testers to design partners. Using methods to procure children’s ideas, input and feedback will contribute the researcher with important insights that could otherwise be hard to obtain. Including children about decisions in their own environment, the products they use and the games they play is a way of recognizing them as actors and participators, rather than onlookers, in our society (Read and MacFarlane, 2006).

3.9 Characteristics of young children to consider when choosing an evaluation method:

- Age
- Verbalization
- Concentration/attention span
- Cognitive skills
- Interests
- Experience in using technologies
- Comfortable and enjoyable test setting
- Knowledge
- Trustworthiness of self-report

4. EVALUATION METHODS FOR CHILDREN’S TECHNOLOGIES

This section gives an overview of existing evaluation methods. Many researchers have looked at how to evaluate technologies with children. All the methods are presented in Table 2 in the appendix.

The emphasis is on methods that include children and are developed for use with children, but expert-methods are also described. Although expert-methods not necessarily include children, the two types of method can complement each other. They can also be used for comparison in order to assess methods. Sometimes the evaluator may have trouble recruiting children or time-pressure, the expert-methods can then come in handy.

4.1 EXPERT/INSPECTION - METHODS

Expert-methods are methods performed by an experienced evaluator. These methods rely on the abilities of the evaluator and do not necessarily include children.

4.1.1 Observation and interview

Observation and interview are traditional evaluation methods. Observation and interview are often a part of other evaluation methods.

4.1.2 Cognitive walkthrough

A cognitive walkthrough is an inspection method where the evaluator steps through the interaction while asking questions regarding how the user reasons and experiences the system.

4.1.3 Heuristic evaluation

Heuristic evaluations often use a set of guidelines developed for (and by) an experienced evaluator. It is common to base an evaluation like this on
one or more existing heuristics (like Nielsen’s HE for example) and combine with concepts from other researchers to adapt the heuristics to the context of the evaluation.

4.1.4 Structured Expert Evaluation Method (SEEM)

The SEEM method is an analytical method developed to assess the usability and fun problems of young children’s computer games. The SEEM method consists of a checklist with questions, originally based on Norman’s theory of action model (Norman, 1998) and on Malone’s concept of fun (Malone, 1980). (Bekker et al., 2010)

4.1.5 (Adapted) Interaction Cycle

The Adapted Interaction Cycle describes how the interaction between a user and a computer game happens in terms of cognitive and physical user actions. Based on Norman’s theory of action model. Developed to be used as a framework for guidelines when evaluating adventure computer games for young children.

4.1.6 Eye-tracking

Looking behaviours are one of the most frequently used measures in studies with infants within developmental psychology (Stephen von Tetzchner, 2010). Preferential looking is a method where an infant is presented with several stimuli. It is measured how long the infant looks at each stimulus and if the average infant looks longer at one stimulus it suggests that the infant can discriminate between the two questions, originally based on Norman’s theory of action model (Norman, 1998) and on Malone’s concept of fun (Malone, 1980). (Bekker et al., 2010)

Franchak et al. used head-mounted eye tracking to study infant’s natural interactions in a playroom. This could be an applicable method in other contexts as well. (Franchak et al., 2010)

4.2 METHODS THAT INVOLVE CHILDREN
4.2.1 Thinking-aloud

Thinking aloud can be used for almost any system. A thinking-aloud test involves having a test subject using a system while continuously thinking out loud. The test-users reveal their view and interpretation of the computer system, and what parts cause the most problems (Nielsen, 1992).

4.2.2 Surveys: questionnaires and interviews

Questioning can take several forms; it can be informal or formal, inclusive or ad hoc, planned or unplanned. When questioning is relatively formal, intended to be inclusive, covers a large number of people, and includes some planning it is often described as a survey (Read, 2008). Surveys have long been used for gathering opinions and information from individuals, and have a history of useful application within HCI and Interaction Design (Read, 2008). The two most common forms of surveys are questionnaires and interview.

4.2.3 Diaries

A diary study can take many forms depending on the goals of the evaluation. The diarists could be the children who are using the product that is being evaluated, or they could be the adults that are observing the children. The ability to describe and reason about abstract concepts and self-reflect is not sufficiently developed in children under the age of 12, who typically can record only more factual information (Markopoulos et al., 2008).

4.2.4 The Fun-Tool kit

The Fun-Tool kit (Read et al., 2002), (Read, 2007), (Read and MacFarlane, 2006) consists of three tools, the Smileyometer, the Fun-sorter, and the Again, again – table. The Smileyometer tool is a Visual Analogue Scale (VAS) based 1 – 5 Likert scale. It is the most used tool in the tool kit so far. It consists of pictorial representations presented in a horizontal row with supporting words. Children are asked to tick one face. If used before trying the technology it can measure
expectations, when used after it can measure experienced feelings or experienced fun during use (Read, 2007).

The Fun-sorter is loosely based on a repertory grid (Read, 2007). It is a tool to compare a set of related technologies or products. It is a cognitively challenging exercise, and for children younger than 8-9 years, each construct should be presented individually (Read et al., 2002). By using different constructs it can measure more than just fun. The items can be presented by picture cards, instead of using words, for children with poor reading or writing abilities.

The Again, again – table provides a lower cognitive load by simply asking children whether they would like to use the product again with the possible options of “yes”, “no” or “maybe”. It makes the child consider each application or product on its own merits, and not comparing several at once. The table can measure endurability and engagement (Read et al., 2002).

4.2.5 Peer tutoring

Peer tutoring is one type of peer collaboration, and means that children teach other children to use the software that is evaluated. The philosophy behind this approach is to view software as part of a child’s play (Höysniemi et al., 2003).

4.2.6 Fun Semantic Differential Scales

The Fun Semantic Differential Scale (FSDS) is an evaluation tool developed for use with young children (Yusrita Mohd Yusoff et al., 2011). The FSDS has been developed through an iterative, participatory-design process including children and staff at a nursery. The tool enables children to express their feelings when interacting with computer products. Pictures of young children showing feelings were used for the scale. In the final design of the scale, the pictures showed four positive emotions (happy, good, love, excited) and four negative (sad, bad, hate, bored), and one neutral option. Labels such as confident and confused were not understood by the children in the previous version of the scale, and excluded.

4.2.7 Wizard of Oz

In a Wizard of Oz evaluation, parts of the interactivity that would normally be controlled by computer technology are imitated or “wizarded” by a human being (Markopoulos et al, 2008). A traditional WOz study has three components: a human wizard, an interface, and a user. During the study the wizard manipulates the interface in such a way that the user is unaware, to varying extents, of the existence and impact of the wizard.

4.2.8 The parent evaluator method

The parent evaluator method is a method for usability testing of interactive products in the field (Markopoulos et al., 2005). The parent act as a facilitator during test sessions executed at home. The parent is requested to maintain a journal where interesting incidents and usage problems must be recorded. The parent at home can also capture data during test sessions.

4.2.9 Contextual laddering

The contextual laddering method refers to a specific one-to-one elicitation interview technique. During the laddering interview a researcher probes into the reasons why children like or dislike a product. The laddering questions can be laddered negative or positive (what did you like or what didn’t you like?). By asking why-questions underlying values can be revealed. This requires abstract thinking, and is cognitively too difficult for very young children. This method can be complimented by other methods, like an observation session or usability test (Zaman, 2008).

4.2.10 Picture Cards method

This method combines the traditional thinking aloud method with picture cards that children can put in a box to indicate problems (Hourcade,
Evaluating serious games for young children

The pictures for the picture cards were chosen from a library of photos for another method, the Picture Exchange Communication System method (PECS). The PECS-method is used to teach non-verbal autistic children to express themselves by exchanging picture cards.

4.2.11 Sorémo

Sorémo is a non-verbal, self-report, affective method that measures children’s emotional state according to nine emotions represented by UI characters (Girard and Johnson, 2009). The method aims to be used in educational software products to investigate the relationship between learning and the emotional states reported. A participatory-design approach was used to construct Sorémo. Children from ages 4-13 years old participated in the design of Sorémo.

5. ASSESSING EVALUATION METHODS (METAMETHODS)

Several papers evaluate usability testing methods by comparing two or more methods. Many researchers compare methods by how many usability problems each method uncovers (Barendregt and Bekker, 2005), (Bekker et al., 2007), (Donker and Markopoulos, 2002). Other approaches include a framework developed by Markopoulos and Bekker, and investigations on the relationship between usability and fun by MacFarlane et al.

5.1 Verbal comments

Van Kesteren et al. assessed six evaluation methods to examine which methods caused more verbal comments from children ages 6-7. They found that active intervention sessions, where researchers asked questions during the tasks, gave most verbal comments. Co-discovery sessions, where pairs of children worked together, did not work as well. Think-aloud, retrospection and peer tutoring worked better (Hourcade, 2007).

5.2 Markopoulos and Bekker - A framework

Markopoulos and Bekker developed a framework to assess usability testing methods with children. When assessing, three dimensions are considered: the criteria to assess the methods, the characteristics describing the methods, and the characteristics of the children being tested. Criteria for assessing the methods: robustness, reliability, validity, thoroughness, and efficiency. Method characteristics: number and grouping of participants, evaluator, context, procedure, data capture, and tasks. The children can be characterized in terms of verbalization, extroversion, gender, concentration, thinking skills, trustworthiness of self-report, knowledge and age (Markopoulos and Bekker, 2003b).

5.3 Usability and fun

MacFarlane et al. investigated the relationship between usability and fun measures (Hourcade, 2007), (MacFarlane et al., 2005). Based on observations of children and children’s own assessment of software they found that there were positive correlations between the two. Scale ratings using a “smileyometer” were not effective, as most children were overly enthusiastic about all the software titles they tried. Ranking the titles based on different characteristics provided more interesting data.

Barendregt et al. have also looked at fun and usability in computer games for children. They state that the quality of a computer game depends both on usability and fun, and have developed a procedure to distinguish usability from fun problems (Hourcade, 2007).

5.4 Bauw et al. – SEEM

Bauw et al. investigated if an expert method could work for children’s technologies by testing a predictive evaluation method. SEEM – Structured Evaluation Method, involves the use of checklists by experts to predict problems in educational games. They found this method could discover most usability problems, but there
were problems that were identified in the expert reviews that did not turn out to be usability problems when children tested the same games (Hourcade, 2007).

6. RESEARCH FINDINGS SUMMARIZED

Findings from research on children’s development, evaluation methods, and serious games are summarized in the table below. The findings can inform the choice and design of an evaluation method. They can function as requirements, and help answer questions such as:

- Which aspects of the game will be evaluated?
- Where should the evaluation take place?
- Who are the participants?
- What is the goal, expected outcome, and final result of the evaluation?
- What methods can be used for evaluating the product?

The table design is inspired by (Höysniemi et al., 2003) and (Markopoulos and Bekker, 2003).

Höysniemi et al. also looked at characteristics of the software to define requirements for the selection of an evaluation method. Markopoulos and Bekker developed a framework to assess usability testing methods with children, using characterizations of children, evaluation methods and assessment criteria.

7. DISCUSSION

Research shows that there is a gap in young children’s evaluation methodologies regarding children under the age of 5. Usability testing with young children is challenging. Methods developed for evaluation with adult users or older children may not be transferable, and researchers and evaluators might not have the needed experience and knowledge about young children. The evaluation methods presented in this paper have been used to evaluate children’s interactive products, computer games, educational games and adventure games. No evaluation methods for serious games in particular were found. Existing methods may be transferable, but needs testing to be validated.

<table>
<thead>
<tr>
<th>Characteristics of young children</th>
<th>Aspects of serious games</th>
<th>Characteristics of evaluation methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Usability</td>
<td>Context: field, lab, evaluator’s office</td>
</tr>
<tr>
<td>Verbalization</td>
<td>Fun</td>
<td>Evaluator alone/team</td>
</tr>
<tr>
<td>Concentration/attention span</td>
<td>Use in real context</td>
<td>Participants</td>
</tr>
<tr>
<td>Cognitive skills</td>
<td>Use over time</td>
<td>Qualitative or quantitative data</td>
</tr>
<tr>
<td>Interests</td>
<td>Effectiveness of the game (e.g. learning)</td>
<td>Experiment or measurement, open or targeted</td>
</tr>
<tr>
<td>Experience in using technologies</td>
<td>Likeability</td>
<td>What is evaluated:</td>
</tr>
<tr>
<td>Comfortable and enjoyable test setting</td>
<td>Children’s opinions</td>
<td>o Usability</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Appeal over time</td>
<td>o Use in real context</td>
</tr>
<tr>
<td>Trustworthiness of self-report</td>
<td>Emotional states</td>
<td>o Opinions</td>
</tr>
<tr>
<td></td>
<td>Age-appropriateness</td>
<td>o Fun problems</td>
</tr>
<tr>
<td></td>
<td>Challenge, curiosity, control, fantasy (Malone and Lepper, 1987)</td>
<td>o Expectations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Measuring fun</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Preference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Likeability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Teach-ability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Learnability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Emotional states</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Use and appeal over time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data capture</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Efficiency and cost</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Validity and reliability of the method</td>
</tr>
</tbody>
</table>

Table 3: Research findings summarized
Although no such methods were found, the characteristics of existing methods may help redesign or design new methods for evaluating serious games.

**The role of children**
While Piaget describes the child as an “onlooker”, newer developmental psychology views the child more as a participant. This is in line with the shift in product design from user-centred design to participatory design, also when it comes to children.

**Expert methods vs. methods involving children**
Expert or inspection methods limit the aspects of the design evaluated to a narrow set of issues captured by the heuristics, guidelines, or questions asked during the walkthrough (Markopoulou et al., 2008). These methods provide little insight into use in a real context, use over time and user preferences. An evaluation involving children (real users), in a real context, will provide a more holistic view of the use and experience of a serious game.

**Lab vs. field**
The lab provides a more controlled environment, but testing in the field means testing in the actual physical and social context where children are likely to use the game evaluated. The developer gets insight into use in unpredictable situations in real life. Testing in the field also places fewer demands on the children to adapt to the testing environment. This can be particularly important for very young children.

**Trade-offs**
When choosing an evaluation method the evaluator needs to decide what is most important for the evaluation. That might involve making trade-offs. For example: What is most important: that children use the tool or method correctly or the quality of the data gained?

**Reliability and validity of methods**
Reliability and validity are important aspects both when choosing an existing evaluation method, designing your own, or assessing an evaluation method. Reliability is concerned with the accuracy of the procedure. The reliability of a method can be assessed be carrying out several trials, and investigate if the results are consistent. Validity is concerned with the evaluation’s success at measuring what the researchers intended it to measure. When choosing an existing method the evaluator should look for papers validating the method. When carrying out a new evaluation method, the evaluator should assess the method regarding reliability and validity.

**Serious games may require new methods**
To be able to assess the effectiveness of a serious game the evaluator needs insight into the use of the game over time and in a realistic context. When the evaluation concerns the purpose and role of a serious game, the game and the participants should be exposed to a realistic usage scenario and context. This might require new evaluation methods.

**Children are not the only users of children’s products**
Children are growing up surrounded by interactive technologies; touch phones, tablets, computers and laptops, gaming consoles and so on. Although children are users of many interactive products, parents have an important role in their use of digital media. They act as gatekeepers, facilitators, teachers and playmates (Shuler, 2009). When evaluating games for children, both in the field and in the lab, the evaluator should have in mind other users and the interaction between the users of the game as well.

**Evaluations are not only about the methods used...**
An evaluation method is just a part of an evaluation plan, and an even smaller part of a design process. Doing an evaluation requires a lot more than just choosing an evaluation method, but it can help the evaluator plan the evaluation. Planning an evaluation ensures that the actual evaluation will run well, and that the results can
be used in a meaningful way. Combining several approaches can help validate results.

**Limitations**
The goals of an evaluation change during the design process. In an iterative design process several evaluation rounds might be needed. This paper does not investigate which evaluation methods are suitable at the different stages of the process.

The paper does not relate the evaluation methods mentioned to children’s developmental theory to a large extent, but that could be an interesting angle for future research.

Only children’s role in the evaluation has been considered in this paper, but the evaluator should have in mind other users as well.

**8. CONCLUSION**

Many researchers have investigated usability testing with children. Nevertheless, there is still a gap in young children’s evaluation methodologies regarding children under the age of 5. Researchers that investigate evaluation methodologies focus on usability and fun, preference and emotional states. Few researchers have investigated appeal and use over time. Both expert-methods and methods involving children have been developed to evaluate children’s interactive products. Some methods have been developed together with children; the Fun Semantic Differential Scale and Sorémo. There are evaluation methods developed for educational games, but no methods for serious games were found.

Findings from the literature search presented in this paper can inform the choice and design of an evaluation (method).

Still, more research is needed to find good and suitable methods for evaluating serious games for young children.

**ACKNOWLEDGEMENTS**

I would like to thank my supervisors Marikken Høiseth and Casparus Boks for valuable guidance and input.
<table>
<thead>
<tr>
<th>Skill/behaviour</th>
<th>Infant Birth - 18 months</th>
<th>Toddler 18 months - 3 years</th>
<th>Preschool 3-5 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention span</td>
<td>Gauged by eye contact.</td>
<td>Gauged by eye contact or involvement with task, bodily movement.</td>
<td>Limited, but increasing. Bright colours, movement are effective.</td>
</tr>
<tr>
<td>Reasoning</td>
<td>Limited to pain and pleasure.</td>
<td>Limited, but concept of 'no' becoming a factor.</td>
<td>Limited, but beginning to be able to know what is liked and what is not.</td>
</tr>
<tr>
<td>Decision making</td>
<td>Does not make complex decisions.</td>
<td>Does not make complex decisions, but 'yes'/no' can be decisive.</td>
<td>Limited, but concepts of what is liked and what is not strengthen. Able to choose one thing over another.</td>
</tr>
<tr>
<td>Understanding</td>
<td>Does not understand scales.</td>
<td>Does not understand scales.</td>
<td>Understanding of simple scales beginning, sorting or identification tasks more effective.</td>
</tr>
<tr>
<td>Motor skills</td>
<td>Possesses some gross motor skills, no fine motor skills.</td>
<td>Rapid gains in gross motor skills, fine motor skills still limited.</td>
<td>Development of both gross and fine motor skills increasing.</td>
</tr>
</tbody>
</table>

*Table 1: Derived from ASTM’s Committee 18 on Sensory Evaluation (Guinard, 2000)*
<table>
<thead>
<tr>
<th>Evaluation methods</th>
<th>What can be tested?</th>
<th>Age</th>
<th>Game/product</th>
<th>Procedure</th>
<th>Context</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert methods - without children present:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walkthrough</td>
<td>Usability</td>
<td>-</td>
<td>Any</td>
<td>Fixed set of questions used</td>
<td>Evaluator’s office</td>
<td></td>
</tr>
<tr>
<td>Heuristic evaluation</td>
<td>Usability</td>
<td>-</td>
<td>Any</td>
<td>Checking if design rules are broken</td>
<td>Evaluator’s office</td>
<td></td>
</tr>
<tr>
<td>Structured Expert Evaluation Method</td>
<td>Usability and fun problems</td>
<td>5 - 7</td>
<td>Adventure games</td>
<td>Walkthrough</td>
<td>Evaluator’s office</td>
<td>(Bauw et al., 2005)</td>
</tr>
<tr>
<td>(Adapted) Interaction Cycle</td>
<td>Usability, help create guidelines</td>
<td>5 - 7</td>
<td>(Adventure) computer games</td>
<td>User action framework</td>
<td>Evaluator’s office</td>
<td>(Barndregt and Bekker, 2004)</td>
</tr>
<tr>
<td>Methods involving children:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>Usability, opinions</td>
<td>-</td>
<td>Any</td>
<td>Questions asked one to one</td>
<td>Lab, field</td>
<td></td>
</tr>
<tr>
<td>Observation</td>
<td>Usability, use in real context</td>
<td>-</td>
<td>Any</td>
<td>Observation of interaction</td>
<td>Lab, field</td>
<td></td>
</tr>
<tr>
<td>Thinking aloud</td>
<td>Usability</td>
<td>7+</td>
<td>Any</td>
<td>Testers verbalize thoughts</td>
<td>Lab</td>
<td>(Markopoulos et al, 2008)</td>
</tr>
<tr>
<td>Surveys</td>
<td>Usability, opinions</td>
<td>-</td>
<td>Any</td>
<td>Questionnaires, rating/ranking scales etc.</td>
<td>Lab, field</td>
<td>(Markopoulos et al, 2008)</td>
</tr>
<tr>
<td>Eye-tracking</td>
<td>Eye movement</td>
<td>14 mths</td>
<td>Toys</td>
<td>Tracking of eye-movements</td>
<td>Lab</td>
<td>(Franchak et al, 2010)</td>
</tr>
<tr>
<td>Diaries</td>
<td>Usability, use and appeal over time</td>
<td>12+</td>
<td>Interactive products</td>
<td>Testers regularly fill in a questionnaire</td>
<td>Field</td>
<td>(Markopoulos et al, 2008)</td>
</tr>
<tr>
<td>The Fun-Tool kit:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Smileyometer</td>
<td>Fun</td>
<td>7 - 9</td>
<td>Children’s technologies</td>
<td>Differential scaling and comparison</td>
<td>Lab, field</td>
<td>(Read, 2007)</td>
</tr>
<tr>
<td>- Fun - sorter</td>
<td>Fun, expectations</td>
<td>7 - 9</td>
<td>Preference</td>
<td>Differential scaling</td>
<td>Lab, field</td>
<td></td>
</tr>
<tr>
<td>- Again, again-table</td>
<td>Endurability and engagement</td>
<td>7 - 9</td>
<td>Computer products</td>
<td>Differential scaling</td>
<td>Lab, field</td>
<td></td>
</tr>
<tr>
<td>Peer tutoring</td>
<td>Real social context, collaboration, teachability, learnability</td>
<td>5 - 9</td>
<td>Computer games</td>
<td>Observation of children tutor and tutee using the system</td>
<td>Lab</td>
<td>(Höysniemi et al., 2003)</td>
</tr>
<tr>
<td>Fun Semantic Differential Scales</td>
<td>Emotional states</td>
<td>3 - 5</td>
<td>Computer products</td>
<td>Differential scaling</td>
<td>Lab, field</td>
<td>(Yusoff et al, 2011)</td>
</tr>
<tr>
<td>Wizard of Oz</td>
<td>Functionality, usability</td>
<td>-</td>
<td>Prototype</td>
<td>Simulation of system</td>
<td>Lab</td>
<td>(Markopoulos et al, 2008)</td>
</tr>
<tr>
<td>The parent evaluator method</td>
<td>Use and appeal over time, realistic context</td>
<td>4 - 6</td>
<td>Interactive products</td>
<td>Field study, diary, interview and observation</td>
<td>Field</td>
<td>(Markopoulos et al, 2005)</td>
</tr>
<tr>
<td>Contextual laddering</td>
<td>Usability, likeability, user experience</td>
<td>5+</td>
<td>Computer games</td>
<td>Interview</td>
<td>Lab, field</td>
<td>(Zaman, 2008, 2010)</td>
</tr>
<tr>
<td>Picture cards method</td>
<td>Usability and fun problems</td>
<td>5+</td>
<td>Computer games</td>
<td>Verbalization/visualization of thoughts</td>
<td>Lab</td>
<td>(Barndregt and Bekker, 2005)</td>
</tr>
<tr>
<td>Sorémo</td>
<td>Emotional states</td>
<td>8 - 11</td>
<td>Educational software products</td>
<td>Differential scaling</td>
<td>Lab, field</td>
<td>(Girard and Johnson, 2009)</td>
</tr>
</tbody>
</table>

**Table 2: Evaluation methods**
REFERENCES


Shuler, C., 2009. iLearn A Content Analysis of the iTunes App Store’s Education Section.


