Conducting Evaluation Studies of Mobile Games with Preschoolers

Laila Shoukry, Stefan Göbel
Multimedia Communication Labs - KOM
TU Darmstadt, Germany
{laila.shoukry, stefan.goebel}@kom.tu-darmstadt.de

Christian Sturm
Hamm-Lippstadt University of Applied Sciences, Germany
christian.sturm@hshl.de

Galal H. Galal-Edeen
Department Of Computer Science & Engineering
The American University in Cairo, Egypt
galal@acm.org

Abstract: In this paper we discuss strategies for evaluating mobile games with three to five year old children with regard to usability and fun aspects. The use of smartphones and tablets have made a lot of interactions of children at this age with technology much more intuitive and made a lot of concerns of previous research of less importance. That said, these devices also pose new usability considerations which have to be addressed. In addition, not all proposed evaluation methods are suitable for evaluating games. As even with careful heuristic evaluations some product-specific problems remain undiscovered until children start using the product, it is better to involve children in evaluations as early as possible. Therefore, we present guidelines compiled from literature and describe our experience during the evaluation phase of our mobile game "Hamza" for teaching preschoolers the Arabic Alphabet.

1 Introduction

Games are the most popular digital activity for children aged two to fourteen, with the highest usage penetration among mobile device users [Gro07]. Digital games fall into a similar category as board games and other self-correcting learning tools and mirror children's natural play interactions like practice play, make-believe play and games with rules [NAE12]. "Digital games have potential as a tool in teaching preschool-aged children because they can provide instant feedback, are flexible, empower children, and foster active learning."(Warren Buckleitner, editor of the Children's Technology Review). Apps are rapidly emerging as a new medium for providing educational content to children. According to a study carried out in 2012 [SLR12], most top-selling paid apps in the education category of the iTunes Store target children and over half of all educational apps target preschoolers. Hamza was designed as a research-based mobile educational game

for Egyptian preschoolers aiming to familiarize preschoolers with the Arabic Alphabet and make them love it in addition to learning letter names and sounds. It uses repetition to help hold learned materials in long-term memory and involves different senses in the learning process by employing attractive audio-visual effects, accelerometer steering and drag and drop. In our design and evaluation, our main priority has been usability and fun, rather than a dense educational content. This is how the story of the game goes: Hamza, the hero of our game has lost his beloved letters and wants to regain them. The letters are drawn as characters with eyes, hands and feet. The player should help Hamza pick all his letters again. This is done using several sub-games (picking the letters while driving a car and dragging the letters into a bag) and then they play with the letters together. All this repetition is meant to teach children, in an indirect way, the names and sounds of the Arabic letters. After completing our first prototype using our proposed Pre-MEGa framework [SSGE12b] as well as surveys of the target population [SSGE12a], it was time to test with target users. Assessing fun and usability in such a product is essential as it will determine if and for how long the children will be using it, which is an important factor for ensuring learning success. Carrying out usability studies with children is not an easy task. The younger the children, the more adaptation or even exclusion of methods used with grown-ups is required. In this paper we will first shed the light on some evaluation methods used with children in literature. As not all evaluation methods will be suitable for our case, we will then narrow down our choices to the most convenient evaluation mechanisms as well as discuss additional aspects encountered during our usability studies.

2 Evaluating Games with Preschoolers

Qualitative evaluation with children can be carried out using observation methods, interviews and/or questionnaires. They usually don't produce (reliable) numeric data but subjective result descriptions. Reviewing literature, it was noted that in most evaluations conducted the researcher either wants to evaluate a certain product on its usability and fun to detect concrete problems and enhance the product or compare different products in terms of usability and fun. Each scenario requires a different approach and uses different testing methods. In observation methods, the evaluator records children's verbal and non-verbal interactions with the product. Depending on the nature of the evaluation, this can be done in many different ways. The first method is simply observing interactions and non-verbal cues [MSH05]. In the "Think-Aloud" Method [Nie94], children are asked to verbalize their thoughts while using the product, which is not always an easy task for children, and if they succeed to keep talking, their remarks are not always reliable. In the "Active Intervention" Method [vKBVL03], the facilitator intervenes by asking the child questions about the experience during the testing session after s/he takes some time familiarizing with the product. The questions are used to prompt the child to verbalize his/her thoughts and usability problems. Another way to do this is the "problem identification picture cards" method [BBB08] where picture cards describing different attitudes towards the product are shown to the child who then chooses the appropriate cards during the test and puts them into a box while eventually explaining the reason behind his/her

choice. This method encourages children to think-aloud and helps them describe their experience. In "peer tutoring" [HHT03], a child first learns how to use a product and is then asked to show another child how to use it. This shows how the first child sees the product and in how far s/he grasps the concepts and instructions. The "constructive interaction" or "co-discovering" [MSH05] approach can be used with products which allow collaboration. Here two or more children use a product together and their natural interactions and conversations are recorded to obtain information about usability, fun and effectiveness. Another creative and playful way to encourage children to think-aloud is the "Mission from Mars" Method [DEI+05] where Children are told that they will communicate with "martians" through voice or video to answer their (occasionally "stupid") questions about the product they are testing. When the usability testing session is video-recorded and then watched together with the child, this is called the "retrospection method" [AWAHAMAN10]. While watching, the researcher asks the child questions about his/her interaction in the video. The "wizard of Oz" Prototyping Method [HHT04] enables testing of early prototypes of gesture-controlled interfaces which are difficult to implement. The idea is to manually simulate the game using simple input methods in the background without testers noticing it and using any other observational methods to record interactions. After interacting with the product to be tested, children can be asked about their opinion and how they rate the product which for adults is usually done through interviews and questionnaires. Adapting these methods to children by reducing the cognitive demand, several methods were proposed like the "smileyometer", the "fun sorter", the "Again/again" [RM06a], the "this or that" [Zam09] and "the drawing intervention" [XRS08] methods. These methods are especially helpful when trying to compare different applications or different activities within an application. A useful framework for evaluating with children is the PLU-E Framework [MR11] which is based on the PLU Model described in [MRMH08]. Using this framework, the evaluator determines a percentage for each of the three variables, P (Playing), L (Learning) & U (Using) to map his product onto the PLU Model and determine the suitable evaluation method accordingly.

Getting verbal data from preschoolers is not an easy task, getting reliable data is another story. It is difficult for children at this young age to properly express themselves in words, they can become shy and are usually inclined to say what they think will please adults [Egl04]. Methods depending on getting children to verbalize their thoughts like the "think aloud" and the "active intervention" were thus found to be unreliable for younger children especially as the interaction with the product already poses a high cognitive load [Zam11].

During usability tests of games, young children don't like to talk while indulged in the game [DR04]. The "Active Intervention" method would distract the children from the game and in our case may make them "lose" in the racing game. In fact, we found that explaining to preschoolers, especially aged three and four, that they are participating in a usability testing, was difficult to interpret and only confusing. To get reliable cues from this delicate age, the experience should be ensured to be as spontaneous as possible: They need to naturally play the game without distractions. Preschoolers also find it difficult to discuss abstract concepts and are still not good at drawing and thus even picture cards and drawing intervention will be difficult. The laddering method was specifically proposed for the age of preschoolers [ZA10] and is especially useful for comparing different products.

However, it doesn't help in detailed detecting of usability problems of a certain game, which was our aim.

Observation, as Kathleen Kremer, the head of the user experience and digital play group at Fisher-Prize explains in [Egl04], is the best method to use when evaluating games with preschoolers. "Peer Tutoring" is also a very good option as preschoolers like to show their friends how they are mastering the game. However, for this method it is best to choose preschoolers who are extrovert and like to use verbal communication as some preschoolers may just show their friends by silently playing the game while their friends watch [HHT03]. During observation sessions the observer makes notes of his/her observations of children's behavior as they interact with the product. The note-taking may be either free-form or based on a previously prepared checklist, depending on the scope of the test and the sample. Here the researcher should ensure children's reactions are as spontaneous as possible by ensuring they feel at-ease. Facial expressions and body language like smiles, laughter, sighs, yawns, frowns, turning away, looking around should be carefully observed as they are generally very good indicators of positive or negative attitudes.

3 Our Approach

The evaluation process of our mobile game "Hamza" consisted of several mechanisms: a field study with preschoolers at the nursery, an online survey for all users of our game were they indicate information about the preschooler using the game and their relation to him/her as well as their personal opinions and evaluations by educators and experts using interviews and survey questions.

Field Study The field study took place at a nursery in Egypt. Pilot evaluation sessions were first conducted with a three year old girl and a four year old boy. We have chosen to use the simple observation method and, to make it more structured, we decided to prepare an evaluation form which we used to take notes during evaluation sessions. The form was prepared using information from literature [BB06], from our own proposed framework Pre-MEGa [SSGE12b] and from our pilot evaluation sessions. This form was especially created for our purpose but depended on general guidelines adapted to the mobile touch screen interface and games as an application type. The evaluation form should be prepared in a way which minimizes the time needed for searching for a certain field or check area. Some items should be repeated for different iterations of playing the game while some are only used in the first time or at the end of a session so it is a good idea to have separate checklists for each iteration and a separate form for each session with all forms for a certain child having background information about the child from the first session on the cover page. The sessions then took place with thirteen children on four consecutive days, where each child left the class for a period of about ten minutes every day to meet with the researcher and play the game. Due to the nature of preschoolers' irregular attendance at nurseries, not all children took part in evaluation sessions on all four days. The children were observed playing the game on two devices: an Android mobile phone and a 7.7 inch

Android tablet and notes were taken by the researcher using the printed form. Based on the child's preference he was allowed to play the game several times. However, when the researcher needed to move on to the next child, the game play was interrupted. For the first session with each child, background information were noted about him/her like name, age and gender, personality characteristics and previous knowledge in the subject matter (in our case the Arabic Alphabet) then the child was asked: "Do you play on the mobile phone or tablet device of your father/mother/both? Or do you own your own device?" (It is also noted e.g. if the child already knows what an iPad is). If the answer is yes to one of these questions, then the next question is "Which game do you usually play?" After answering the questions, the child started playing the game without getting any instructions or explanations (The children usually did not have enough patience to even answer any questions and wanted to start playing once they saw the mobile device). This is because a mobile game for preschoolers should be self-explanatory. During game play the observer didn't intervene and silently took notes. Occasionally, we have also used peer tutoring with preschoolers who showed good verbal skills and quickly mastered the game. For evaluating the difficulty level of the game it is a good idea to especially consider the youngest age to be examined because at the age of preschoolers drastic advances in skills can take place in just a few months.

Surveying the Parents Parents' comments and suggestions not only help improve the design of a learning game targeting children but can also give valuable feedback on its educational effectiveness and on what additional educational needs designers need to address in future updates or new games for this age group. As the first version of our game Hamza was already available for free download on Google Play, this greatly facilitated our effort to collect a large set of data. For this we have placed a link at the end of the game asking users to fill our online survey. In two months we had over 90 filled surveys which helped us to make several updates to the game during this time period.

4 Results

Field Study The author of [RM06b] discourages applying statistical tests to children's responses as they are affected by a lot of different aspects and can hardly be generalized. She suggests that the researcher try to look for trends and outliers, as usability tests with small groups of children often give only a general feel for the product to be tested. From the qualitative data gathered using our evaluation forms, gender and age differences in preference and skills could be extracted as well as specific usability problems in the game. The following are some results of the usability testing sessions: Overall the children seemed to enjoy the game. Most of them asked to play it again in the same sessions and some of them asked to play it for the fifth time and had to be interrupted by the researcher. Most children needed no or only minimal help and could get along on their own. Girls were found to ask for help more than boys, especially when they have several choices to choose from. The racing game was difficult for some 3-year-olds but most of them mastered it at the second or third time. The drag and drop game was suitable for all ages but

only one or two children needed instructions in the first time on how to do it. Boys seemed to prefer the racing game and were better at steering by tilting the device. Girls seemed to prefer the drag and drop game and some of them had some difficulties steering the car. Girls seemed to enjoy the songs whereas most boys would just skip them to play games. In the navigation of version 1 of the game, it was not easy to repeat a sub-game without starting the game from the beginning, so this had to be improved. The alphabet menu was not intuitive for most children and they would need an adult to choose the suitable level for them. Most children seemed very excited about the positive feedback they get when they "eat" an alphabet in the racing game and some started to say comments like: "I caught it!".Although we had expected that the button which leads to the online evaluation at the end of the game might cause a usability problems as children might click on it and then go out of the game, this didn't happen in any evaluation session of the game. This might be due to the fact that the home button on the same screen was more catchy to the children due to the picture than the text which was beneath for parents to read. This showed that this screen needed no further altering. The short ranges of the alphabet were found to be a successful idea as children didn't get bored due to the short sub-games. However, when they directly clicked on the car on the alphabet choice screen they automatically chose to include all alphabets in the game which sometimes made the sub-games a little boring for them because they were much longer. Some buttons were found difficult to click on like the pause button in the racing game so it had to be enlarged. The area for dragging and for dropping the alphabets into the bag had to be increased as it was difficult for some younger children to do the dragging seamlessly. Drag and drop on the smaller mobile device was more difficult than on the tablet device. Racing was much easier on the smaller mobile device than on the tablet device because steering needed little effort and balance. Most children understood the sign for pause and play buttons and good get along very well on their own without needing additional help. Most children preferred to play using the tablet device upon seeing both devices. The racing game speed was still fast for some children, especially younger children.

Parent Survey The surveys showed very positive ratings of different aspects of the game as well as how people described their children's use and benefit of the game. Additionally, the surveys revealed the following dependencies:

The following suggestions were the most important ones which we have considered in versions 2 and 3 of the game: 1) Using more repetition of the letters sounds during the racing game, 2) making the letter names be pronounced in formal Arabic instead of Egyptian Arabic, 3) adding more episodes to reinforce learning, 4) making voice more clear and enhancing navigation.

Game Updates Based on all evaluations carried out, a new task list was created for version 2 of the game. Version 2 was released after one week containing a lot of updates based on users' remarks and usability test results. Version 3 was released after a month with further updates. The following were some updates in further versions of the game: 1) Enhancing navigation to allow replaying sub-games without repeating gender and alphabet choices and skipping all subgames and animations. 2) Enlarging buttons which

were difficult to hit. Increasing the area for dragging and for dropping the alphabets into the bag. 3) Reducing the speed of the racing game and the tilt reaction. Enhancing audio quality. Adding a bag test game to reinforce learning. Adding more instruction for bag game and animating the bag saying the instructions (v. 2 and 3). 4) Using more repetition of the letters sounds during the racing game (v.3). 5) Making the letter names be pronounced in formal Arabic instead of Egyptian Arabic (v.3).

5 Conclusion and Future Work

In this paper, we have described the evaluation process of the Android game Hamza teaching preschoolers the Arabic Alphabet. The best way for evaluating products with this young age was found to be the observation method guided by an evaluation form designed apriori where the observer can take notes of different behaviors and interactions. Hamza Game is now on Google Play and has over 98 thousand users and over 360 positive ratings with an average rating of 4.4/5. Due to the success of the game we are now developing Hamza2 helping acquire second-level literacy skills. We are also working on producing Hamza1 in different languages.

References

[AWAHAMAN10] Areej Al-Wabil, Luluah Al-Husian, Rana Al-Murshad, and Abeer Al-Nafjan. Applying the retrospective think-aloud protocol in usability evaluations with children: Seeing through children's eyes. pages 98–103. IEEE, 2010.

[BB06] Wolmet Barendregt and M. M. Bekker. Developing a coding scheme for detecting usability and fun problems in computer games for young children. *Behavior research methods*, 38(3):382–389, 2006.

[BBB08] Wolmet Barendregt, Mathilde M. Bekker, and Ester Baauw. Development and evaluation of the problem identification picture cards method. *Cognition, Technology & Work*, 10(2):95–105, 2008.

[DEI+05] Christian Dindler, Eva Eriksson, Ole Sejer Iversen, Andreas Lykke-Olesen, and Martin Ludvigsen. Mission from Mars: a method for exploring user requirements for children in a narrative space. pages 40–47. ACM, 2005.

[DR04] Afke Donker and Pieter Reitsma. Usability testing with young children. pages 43–48. ACM, 2004.

[Egl04] Tammie Hutto Egloff. Edutainment: a case study of interactive cd-rom playsets. *Computers in Entertainment (CIE)*, 2(1):13–13, 2004.

[Gro07] NPD Group. Amount of time kids spend playing video games is on the rise., 2007.

[HHT03] Johanna Höysniemi, Perttu Hämäläinen, and Laura Turkki. Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interacting with Computers*, 15(2):203–225, 2003.

[HHT04] Johanna Höysniemi, Perttu Hämäläinen, and Laura Turkki. Wizard of Oz prototyping of computer vision based action games for children. pages 27–34. ACM, 2004.

[MR11] Lorna McKnight and Janet C. Read. Plu-e: a proposed framework for planning and conducting evaluation studies with children. pages 126–131. British Computer Society, 2011.

[MRMH08] Panos Markopoulos, Janet C. Read, Stuart MacFarlane, and Johanna Hoysniemi. *Evaluating children's interactive products: principles and practices for interaction designers*. Morgan Kaufmann, 2008.

[MSH05] Stuart MacFarlane, Gavin Sim, and Matthew Horton. Assessing usability and fun in educational software. pages 103–109. ACM, 2005.

[NAE12] NAEYC. Technology and interactive media as tools in early childhood programs serving children from birth through age 8. A joint position statement., 2012.

[Nie94] Jakob Nielsen. *Usability engineering*. Elsevier, 1994.

[RM06a] Janet C. Read and Stuart MacFarlane. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. pages 81–88. ACM, 2006.

[RM06b] Janet C Read and Stuart MacFarlane. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 conference on Interaction design and children*, pages 81–88. ACM, 2006.

[SLR12] Carly Shuler, Zachary Levine, and Jinny Ree. iLearn II: An analysis of the education category of Apples app store. In *New York: The Joan Ganz Cooney Center at Sesame Workshop*, 2012.

[SSGE12a] Laila Shoukry, Christian Sturm, and Galal H. Galal-Edeen. Arab Preschoolers, Interactive Media and Early Literacy Development. In *The International Conference on E-Learning and E-Technologies in Education (ICEEE)*, pages 43–48, September 2012.

[SSGE12b] Laila Shoukry, Christian Sturm, and Galal H. Galal-Edeen. Pre-MEGa: A Proposed Framework for the Design and Evaluation of Preschoolers' Mobile Educational Games. In *The International Conference on Engineering Education, Instructional Technology, Assessment, and E-learning*. Springer, 2012.

[vKBVL03] Ilse EH van Kesteren, Mathilde M. Bekker, Arnold POS Vermeeren, and Peter A. Lloyd. Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects. pages 41–49. ACM, 2003.

[XRS08] Diana Xu, Janet C. Read, and Robert Sheehan. In search of tangible magic. pages 97–100. British Computer Society, 2008.

[ZA10] Bieke Zaman and Vero Vanden Abeele. Laddering with young children in User eXperience evaluations: theoretical groundings and a practical case. pages 156–165. ACM, 2010.

[Zam09] Bieke Zaman. Introducing a pairwise comparison scale for UX evaluations with preschoolers. pages 634–637. Springer, 2009.

[Zam11] Bieke Zaman. Laddering method with preschoolers. Understanding preschoolers' user experience with digital media. 2011.