Understanding the Fidelity Effect when Evaluating Games with Children

Gavin Sim
ChiCI Group
University of Central Lancashire
Preston, UK.
44 1772 895162
grsim@uclan.ac.uk

Brendan Cassidy
ChiCI Group
University of Central Lancashire
Preston, UK.
44 1772 893265
bcassidy1@uclan.ac.uk

Janet C Read
ChiCI Group
University of Central Lancashire
Preston, UK.
44 1772 893285
jcread@uclan.ac.uk

ABSTRACT
There have been a number of studies that have compared evaluation results from prototypes of different fidelities but very few of these are with children. This paper reports a comparative study of three prototypes ranging from low fidelity to high fidelity within the context of mobile games, using a between subject design with 37 participants aged 7 to 9. The children played a matching game on either an iPad, a paper prototype using screen shots of the actual game or a sketched version. Observational data was captured to establish the usability problems, and two tools from the Fun Toolkit were used to measure user experience. The results showed that there was little difference for user experience between the three prototypes and very few usability problems were unique to a specific prototype. The contribution of this paper is that children using low-fidelity prototypes can effectively evaluate games of this genre and style.

Categories and Subject Descriptors
H.5.2 Information interfaces and presentation (e.g., HCI): User Interfaces - Evaluation/methodology, Prototyping.

General Terms
Human Factors; Design; Measurement.

Keywords
Prototyping; Evaluation; User Experience; Usability; Children

1. INTRODUCTION
The game industry is a multi-billion dollar concern, with games being developed for a variety of devices and emerging technologies. There are financial pressures to ensure the rapid development of games, ensuring that the game gets to market and is differentiated from its competitors. To ensure games are successful it is extremely important to playtest them as early, and as often as possible during the development. This is necessary to improve the usability, and address game balancing and motivation issues [1].

IDC '13, June 24 - 27 2013, New York, NY, USA

Without this feedback, the player experience may not be optimal and players may switch to an alternative game. Usually user experience is evaluated after there is a working prototype implemented which it is ready for beta testing [2]. However, in the early stages of development prototypes can take the form of game sketches and thus, for some testing, a fully functional prototype may not be necessary. Time constraints and budgetary limitations often influence the fidelity of the prototype being developed.

Prototypes are developed in a number of forms and two distinct categorizations of prototypes have been identified, these being low and high fidelity. The development of low-fidelity prototypes is usually associated with the use of material different from the final product, such as paper sketches [3]. The aim of these early sketches are to open the design space for new alternatives [4]. In contrast high-fidelity prototypes usually offer a level of functional interactivity using materials that you would expect to find in the final product, for example a smartphone touch screen [3].

In order to evaluate a prototype there are two main evaluation methods: inspection based and user testing. The most widely researched inspection method is the heuristic evaluation method developed by Nielsen and Molich [5]; in recent years bespoke heuristic sets have emerged for evaluating games [6-8]. In user testing, people from the target user group interact with the prototype or product. During this interaction, their behavior and experiences are collected using a variety of techniques including observations [9] and think aloud [10]. However, when evaluating prototypes the results can be influenced by the fidelity effect associated with the form of the prototype. In a study examining usability, many of the problems were not reported in the low-fidelity version [11] as they were associated with the functionality of the device. In another study it was concluded that users appeared to over compensate for deficiencies in aesthetics in low-fidelity prototypes [12]. When evaluating prototypes of games designed for children, understanding the effect fidelity has on the results, is clearly desirable.

When a game is aimed at children, user testing is a credible option to evaluate usability and user experience. However, many traditional adult evaluation methods are ineffective when used with children [39] and so adaptations to evaluation methods are necessary. The behavior of the evaluator may affect the children's performance, as might other factors, including the decoration of the room and the observational equipment being deployed [13].

Within the area of Child Computer Interaction (CCI) a great deal of early work focused upon the inclusion of children in the design process both as developers of prototype applications [14] and as

users in usability studies [15]. These studies highlighted that, given an appropriate method, children can successfully design and evaluate a range of technology, software and products. There has been very little research analyzing the fidelity effect of prototypes with children, especially within the context of games; this formed the motivations for this research study.

This paper is structured as follows; first, there will be an analysis of the existing body of knowledge relating to the fidelity effect and evaluation methods with children and this analysis will lead to a number of research questions. Then the research method will be presented along side the findings from this study. Finally a discussion of the results is presented with the implications for designers and developers of games.

2. Related Work

Literature from two domains informed this work; research on comparing the fidelity effect with prototypes and research on the use of evaluation methods with children.

2.1 Prototype Fidelity

Prototypes are used for a variety of different purposes including the evaluation of design ideas, exploration of ideas and as props to assist with communication as part of the development process [16, 17]. Despite the advantages of using a low-fidelity prototype in terms of costs, care needs to be taken in the validity of any findings from evaluations, as the fidelity of the prototype can influence the results obtained. When using prototypes it can be difficult to always understand if the findings from evaluations are closely aligned to the concept of the artifact being evaluated or are more associated with the characteristics of the prototype itself [18]. There have been many researchers who have discussed the merits of different prototype techniques, including [3, 19], but many claims have not been validated based on empirical evidence and there are contradictions in the literature. For example it was claimed that *When something appears to be finished, minor flaws stand out and will catch the users' attention.* [19], however a study looking into the visuals for game prototypes disputed this claim [20]. This study, [20], evaluated a serious game prototype and the majority of participants who were exposed to either the low-fidelity or high-fidelity prototype never referred to the visuals. This study focused specifically on usability problems, whilst if user's emotional responses are sought, it is claimed that developers tend to use higher-fidelity prototypes characterized by considerable aesthetic refinement [12]. This leads to the issue of determining which fidelity of prototype to use when the purpose of the evaluation may be to capture both usability problems and emotional responses, which is often necessary within the context of games.

There have been a number of comparative research studies of prototypes at different fidelity levels [18, 21, 22], however the results are rather inconclusive. There have been studies showing that, with low-fidelity prototypes, results can be gathered that are equivalent to those gained from evaluating fully operational products as well as other studies reporting the additional benefits of higher fidelity prototypes [23]. Seflin *et al.* [21] investigated whether subjects confronted with a paper-based low-fidelity prototype differ in their willingness to criticize a system, compared to a computer based prototype. The results showed that there was no difference in the number of criticisms but the users preferred the computer prototype.

A concern in the context of games designed for children is that the majority of comparative studies have been performed with adult users, and therefore it is unclear whether the findings can be generalized to children.

There are very few research studies that have looked at the fidelity effect when evaluating prototypes with children. In a study using 16-17 year olds comparing low and high fidelity prototypes for tabletop surfaces [24], the findings suggest that one should be cautious in generalizing high-level user interactions from a low-fidelity prototype towards a high fidelity prototype. For example it was feasible to layer information on top of each other in a 3D space and this was not feasible in the 2D space. Within the context of games and mobile interaction the device could clearly influence the results, as the interaction could not be simulated in one context. This has also been reported in a study by [20] with one participant struggling to use the accelerometer being only able to use it along one axis.

Within the context of games there is very little research on the fidelity effect with children. In a study looking at visuals for games aimed at children between 4 and 7 the evaluators were adults aged 20-28 [20]. This study showed that usability problems could be found irrespective of prototype fidelity. However, there is concern over the results as they were not derived from user testing (using participants aged 4-7) thus the data may contain false positives which may affect the validity of the conclusions.

Therefore it is necessary to understand the fidelity effect to establish whether the same findings occur when the participants are children.

2.2 Children Evaluating Technology

When evaluating technology with children it is important to clearly establish the purpose of the evaluation and clearly understand the data that is to be captured. Evaluation methods tend to focus on either usability or user experience. However, the emphasis over the last few years has moved away from usability evaluation to focusing on the user experience. It has been suggested that user experience is not clearly defined or well understood within the HCI community [25] and the CCI community. When compared to traditional usability, user experience differs significantly in the constructs that are measured. ISO define user experience as *a person's perception and responses that result from the use and/or anticipated use of a product, system or service [26]*. Usability evaluations tend to focus on task performance whereas user experience focuses on lived experiences [27]. User experience is subjective and therefore cannot be captured using traditional usability metrics like task completion time or error rates. User experiences that can be captured can include physical, sensual, emotional and aesthetic experiences: for example, if the objective of the evaluation were to measure fun, then metrics would be required to capture these emotions.

Carroll suggested that things are fun when they attract, capture, and hold our attention by provoking new or unusual emotions in contexts that typically arouse none [28]. Fun is one attribute of user experience that is important to measure as it is one of the major motivations for children to interact with technology [29] and one of the important factors associated with games. Malone pioneered the study of fun as an important aspect of software for children [30]. Without the technology providing a positive experience, children are unlikely to interact or accept it. Therefore fun is an important construct to measure as part of a user experience study with children.

Within the context of user experience, several evaluation methods have emerged for use with children including Problem Identification Picture Cards [31], the Fun Toolkit [32] and Laddering [33]. Many of these new methods for evaluating user experience rely on the use of survey instruments or techniques. The use of survey methods with children often brings into question the validity and reliability of children's responses [34]. This is in part due to the large differences in cognitive and developmental abilities between children of the same age [35]. This can lead to well known issues such as satisficing, suggestibility and misunderstanding [36]. Maximizing the reliability of children's responses is vital to ensure the validity and integrity of results and to give strength to any subsequent design recommendations or decisions.

Whilst these user experience methods have all been tested and validated with children this has mainly been done in isolation of other methods or against more traditional survey methods designed for use with adults. Research was conducted which compared the Fun Toolkit with the This or That method when evaluating games [37]. The results of this study showed that the two methods yielded very similar results and were comparable for identifying game preference. This is important because if the alternative methods were to yield different results, there is the potential for the results of entire research studies to be questioned and possibly be flawed. This could in turn be very costly both in time and money, especially if the design of an application or piece of technology has been based on these results.

For usability testing with children, researchers have examined think aloud, interviews and the use of questionnaires [15]. It has been shown that children can identify and report usability problems. For example children were able to detect usability problems which would aid the design of a physically and vocally interactive computer game for children aged 4-9 [38]. However when conducting usability research with children there are a number of challenges that need to be considered. In one study [39] out of 70 children only 28 of them made verbal remarks during the user test. This may well be attributed to their personality, a study showed that personality characteristics influences the number of identified problems [40], therefore further research is still required to understand usability methods and their limitations and to ascertain which are applicable to children. One area worth considering is inspection methods as these have largely been neglected.

Given the fact that children can report usability problems, and there are valid tools for measuring user experience with children, it should be possible to evaluate the fidelity effect of prototypes with children. Therefore this raised three questions when evaluating prototypes with children:

1. Would the initial expectations of a game be lower for children who are presented with a low-fidelity prototype compared to the higher fidelity prototype?
2. Would there be any difference between children's overall rating of a game depending on fidelity?
3. Are different usability problems reported depending on fidelity?

3. Method

This study used a between subject design, in which the user experience of a single game was evaluated in three different fidelities. Each child would either play a low-fidelity sketched version of the game, a mid-fidelity version based on screen shots of the actual game or a high-fidelity functional game.

3.1 The Game Prototypes

The Farm Match game for the iPad3 was selected for use in this study, as it would enable the game to be reverse engineered into two lower fidelity prototypes that would be playable by the children, see Figure 1.

Figure 1. Left image is the sketch of the original user interface on the right

The decision was made to use a game that would be interactive and playable in a low-fidelity form, as a number of games developed for the iPad were originally board games or card games such as Monopoly and Solitaire and therefore would offer similar levels of interaction. Other games such as Angry Birds were considered, but simulating the gameplay within a paper prototype was judged to be potentially problematic, as the interactivity and animations could not easily be simulated.

This study was designed to establish, given prototypes exhibiting the same levels of functionality, whether children's reported experiences are similar. Two prototypes of the original farm match game were created; for one prototype (screen), screen shots of the game were captured and printed out in colour on A4 paper. The interface components, such as, the menus and cards containing the animals which the children could turn over and interact with, were individually cut out and placed on top of the background screen. Using the screen version as a template, a second, lower fidelity, prototype (sketch) was created by sketching each of the screens. These two prototypes (sketch and screen) each consisted of 4 screens of the game, each on a separate piece of A4 paper. The facilitator was responsible for moving between the pages depending on the users' selection. Within the game, the child could turn over the individual pieces of paper or post-it notes, to view the images and to establish whether two items matched. If they did not match the children were instructed to turn them over ensuring they were in the original position.

Despite the fact that low-fidelity prototypes are usually quite different in form and function from their final version, the decision to reverse engineer the game was based upon the fact that it isolates fidelity from maturity of design, which is important to reduce confounds in this study.

3.2 Study Design

The study aimed to establish whether the fidelity of a prototype of a matching game designed for children affected user experience and usability.

There are numerous evaluation methods that could be adopted for measuring user experiences, however, it is important that the methods have been validated with children, and therefore some elements of the Fun Toolkit were selected. This tool has predominantly been used for comparative analysis of technology or games with children. This study was a between subject design and therefore the Fun Sorter was omitted as this required the children to compare one experience with another. Thus the Smileyometer and the Again Again tools were used in this study.

The first tool is the Smileyometer, this is a visual analogue scale with the coding based upon a 5 point Likert Scale, with 1 relating to 'Awful' and 5 to 'Brilliant' (see Figure 2).

The Smileyometer is usually used before and after the children interact with the technology. The rationale in using it before is that it can measure their expectations, whilst using it afterwards it is assumed that the child is reporting experienced fun. The Smileyometer has been widely adopted and applied in research studies [37] to measure satisfaction and fun as it is easy to complete and requires no writing on behalf of the children.

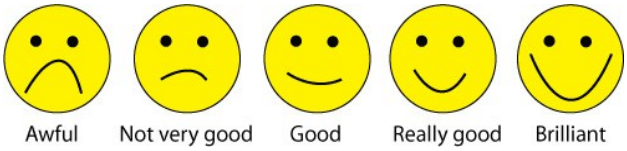

Figure 2. Smileyometer rating scale

The Again Again tool requires the children to pick 'yes', 'maybe' or 'no' for each activity they have experienced. In this study the children were asked 'Would you like to play this game again?' and they had to respond accordingly. An example of the completed Again Again table can be seen in Figure 3.

Figure 3. Completed Again Again table

There are many methods available for evaluating the usability of software for children. However there is very little research on the effectiveness of inspection methods with children acting as expert evaluators. Therefore the decision was made to use an adaption of the cognitive walkthrough [41] by incorporating observational techniques. In a cognitive walkthrough, expert evaluators interact with the technology and their actions and responses are evaluated according to the users' goals through a series of questions related to the cognitive model. Problems are identified through differences in the evaluators' expectations and the steps required to complete the action. The evaluators usually document and report these differences, however for this study observers were used to capture and document issues, rather than the children having to report or verbalize their actions. In this study a number of action sequences within the game were identified, such as starting the game and returning to the home page, with the overall goal to play the game. These actions were then used to formulate a number of tasks which were then observed whilst the child interacted with the prototype. If the child performed the wrong action the observer would document this. Using observers reduced the demands of the cognitive walkthrough, as the children did not need to personally document the issues they encountered.

3.3 Participants

The participants in this study were 37 primary school children from a UK school; the children were aged 7-9 years old. They took part in this study during their normal school day, and there was one researcher for each prototype version. The three researchers who took part in this study all had experience working with, and conducting evaluations with children of these age groups.

3.4 Apparatus

The researcher gave the children a pen and a data capture form to complete the Smileyometers and Again Again table. The researchers captured the responses to the tasks and noted any usability problems on a separate sheet, noting down the child who was responding, so this could be matched with the data from the Fun Toolkit.

3.5 Procedure

The research was conducted in an empty classroom within the school. As the study was a between subject design, the children were required to play with either the Sketch, Screen or iPad version of the game. The children came in groups of 3 and were allocated to one of three separate desks. On two of the desks there was a prototype of the game being evaluated (sketch or screen), the third desk simply had an iPad. On the two desks with the paper prototypes there was also an iPad with the game loaded so these children could play the real game once they had finished evaluating the prototype versions.

Before the child played the game, four initial questions were asked about their age and experience of playing games. Following this, the first screen of the interface was shown to them based upon the evaluation condition, and the child was asked to complete the first Smileyometer to measure their expectations of the game before playing.

The children then played the game / prototype and were asked a number of questions, which acted as play prompts. These were;

- How do you think you start the game?
- Select the game option with 8 tiles.
- Play the game by turning over the two cards. If they match leave them turned over. If not, return to their original state.
- Where do you think the score is located?
- How do you think you would play again?
- How would you go back to the start page from the game?

For the two low-fidelity versions of the game, the researcher turned over the pages to simulate the interaction and asked the appropriate questions or instructed the children on the next task. For each of the questions or tasks the children's actions or response were recorded by the researcher on a data capture form. If the children encountered any usability problems whilst playing the game these were also documented.

Once the child had played the game once, the child was asked to complete the Again Again table and the second Smileyometer.

Approximately 5-10 minutes was allocated to play each game, but this was flexible to allow children to stop earlier if they were bored, or continue longer if they were engaged. For ethical reasons, all the children who evaluated the lower fidelity prototypes (sketch and screen versions) had the opportunity to play the iPad version of the game, once the study had concluded. Thus, for each child the study lasted between 15-20 minutes.

3.6 Analysis

All children managed to complete the Smileyometers before and after they played the games. They were coded in an ordinal way 1-5, where for example 5 represented 'brilliant' and 1 'awful'. The Again and Again table, resulted in a score for each game with yes being coded as 2, maybe 1 and no 0.

For the set of questions that were used as play prompts, findings relating to usability were calculated based on whether a child had an issue or not for each particular task/question. For example, if the child pressed the wrong option to start the game this would be recorded as a single usability issue and logged as a frequency 1 - if a second child had the same difficulty, the frequency would rise to 2. The decision was made to only count problems once irrespective of whether an individual selected the wrong option multiple times, as this measurement was concerned with the number of children who encountered difficulties within a particular task. In addition the problems reported, within each prototype, were catalogued and merged using an open card sort to produce, for each prototype, a consolidated list of usability problems. The 3 lists of problems were then compared to establish if any specific usability problems were fidelity related.

During this analysis of the three lists of problems one problem that was reported on the iPad was deemed a false positive. The observer noted an issue that the child was popping balloons, however this was judged not to be a real problem, as, when the game finishes balloons appear and it is possible to pop these, see Figure 4.

Figure 4. Screen indicating the game has finished

Popping the balloons is not actually a problem as it does not affect the overall game, and it is just an animation with a level of interactivity that indicates the game has finished. Therefore the decision was made to remove this from the final problem set for the iPad.

4. Results

The results for the Fun Toolkit are initially presented followed by the analysis of the usability problems associated with each prototype.

4.1 Fun Toolkit

Each of the 37 children completed the Smileyometer before and after they played the game and the results for each prototype are presented in Table 1.

Table 1. Mean Scores and Standard Deviations for the Smileyometer

Prototype	Before		After	
	Mean	SD	Mean	SD
Sketch	3.42	.793	4.25	.965
Screen	3.75	1.055	3.75	.965
iPad	3.62	.870	4.31	1.032

Before the children had played the game, the Smileyometer results suggest that the children anticipated that the game would be between good and really good. The screen version of the game had the highest mean score before the children had played the game and the lowest after. For the other two versions of the game (Sketch and iPad versions) the mean scores increased after the children had played the game, suggesting that the game had surpassed their initial expectations. The mean scores for the Sketch and iPad versions after the children had played the game were very similar.

The Again Again table that forms part of the Fun Toolkit was also used to establish if the children wished to play the game again and the results for this are shown in table 2 below.

Table 2. Frequency responses to whether a child would play it again

Prototype	Yes	Maybe	No
Sketch	7	3	2
Screen	8	3	1
iPad	9	4	0

It is clear that the majority of children in all 3 conditions would like to play the game again. As expected, none of the children who played the iPad version indicated they did not wish to play it again.

4.2 Usability Issues

The number of children who encountered usability problems for each of the specific tasks is shown in table 3 below.

Table 3. Number of children who had usability problems for each of the tasks

	Number of Problems Reported		
Task	**iPad**	**Screen**	**Sketch**
Start game	8	5	4
Select 8 tiles	1	2	4
Playing game	0	0	2
Score located	1	1	3
Play again	1	1	3
Return home	0	3	2
TOTAL	11	12	18

Despite the fact that there appeared to be more problems in the sketched prototype, these problems were all attributed to 7 of the 12 participants. For the iPad and Screen versions, 10 children in each condition encountered difficulties in completing all the tasks successfully.

Following an open-card sort the individual problems (many of which were the same) were merged within fidelity and the number of distinct reported problems for each prototype is shown in table 4.

Table 4. Number of distinct usability problems for each Prototype

	iPad	**Screen**	**Sketch**
Number of distinct problems	4	5	7

A comparative analysis between these distinct problems showed there were similarities in the problems identified, with many problems being reported in all three versions. For example the children had problems in starting the game in all versions, the children selected the *Full Game* button instead of the Paw. The problems were therefore merged into a consolidated list and in total there were 7 unique usability problems observed, figure 5 shows the number of problems reported in each of the three prototypes.

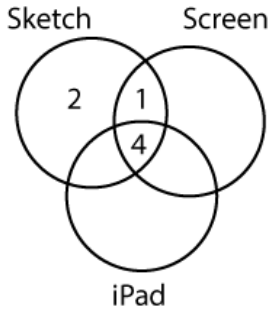

Figure 5. Usability problems in each prototype

There were four problems identified across all three prototypes these were:

- Unable to start the game
- Selected the wrong number of tiles
- Unable to locate the score on the screen
- Unable to identify which button to press to replay the game

One problem was only identified in the sketch and screen version, this was:

- Unable to identify the button to return to the home page

In total only 2 problems were identified that were specific to a single fidelity; both were in the sketched version:

- Unsure how to turn the cards over
- Didn't recognize matched items

5. Discussion

In this study an iPad game was reverse engineered to construct two prototypes at a lower fidelity. Our study aimed to understand the impact fidelity would have when evaluating user experience and usability with children. Three research questions were identified through an analysis of the literature.

The first question aimed to discover whether the initial expectations of children would be lower when presented with a low-fidelity prototype. The mean scores before the children played the game were similar with the Screen version having the highest mean. It was expected that the iPad version would have the highest score but this did not prove to be the case. In a study examining aesthetics in prototypes with adults, users appeared to compensate for deficiencies in aesthetic design by overrating the aesthetic qualities of reduced fidelity prototypes [12]. The issue of overcompensation might have occurred with the children being over enthusiastic when rating, for example the screen version, when rating the game with the Smileyometer; additionally it has been shown in other studies that children are generous in their evaluations of software [42], but this is generally associated with younger children. Although the age range of the children was 7-9 these were balanced within each groups, see Table 5 for average age of the children in each group.

Table 5. Average age of children in each group

	iPad	Screen	Sketch
Average age in years	8.08	8.16	8.15

The second research question aimed to establish if there is any difference, depending on fidelity, between children's ratings of their overall experience of playing the game. The results of the Smileyometer after the children had played the game suggest that the low-fidelity sketch version is similar to the high-fidelity iPad version, whilst the screen version was lower. However the Again Again table showed that only 58% of the children who played the sketched version stated that they wished to play the game again. This is compared to 69% for the iPad and 75% for the screen version. All three versions were favored by the children in so far as there were no children stating that they did not want to play the game again on the iPad, only one did not wish to play on the

screen and just two did not want to play on the sketch version. It would appear as though the game experience can be predicted through analyzing lower fidelity prototypes (within the constraints of this game genre) using the two tools within the Fun Toolkit.

Finally, clarity was sought as to whether different usability problems are reported depending on fidelity. In total 7 different problems were reported and only 2 of these were unique to a specific prototype. The two problems were unique to the sketched version raising questions as to how and why this should be.

The first unique problem was the fact that a child was unsure how to turn the cards over. Each drawing was on a post-it note and placed face down on the paper, the child just needed to turn over the post-it note to reveal the image. This would not be a problem in the iPad version as the device automates this process, however, the issue was easily rectified by intervention from the facilitator.

The second problem was that one child did not recognize that the items were matched. For this study the items were drawn separately on post-it notes and they were judged by the researcher to be similar and easily identifiable, see Figure 6. It might have been worthwhile if only one drawing was made and then photocopied.

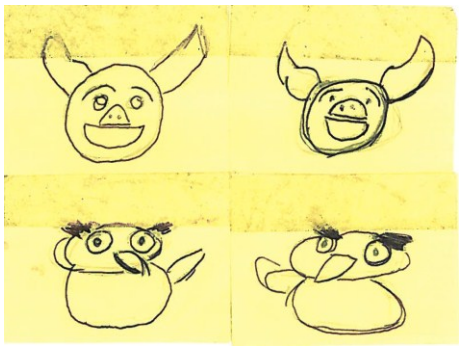

Figure 6. Sketched versions of the tiles

Overall it would appear that very few problems are unique to a specific prototype. Within the context of games that can be simulated on a paper prototype, if care is taken in the construction of the low-fidelity prototype then this should reveal similar usability problems as the higher fidelity versions.

6. Conclusion

This study showed that the children reported similar experiences before and after game play, in the three prototypes. Therefore for games where the interaction and game play can be simulated on paper, it is possible for games developers to successfully evaluate games concepts, design ideas and interaction with children at early stages of games development using paper prototypes.

With regards to usability, problems were identified using an adaptation of the cognitive walkthrough incorporating observational techniques and the results suggest that this may be a viable technique to identify usability problems with children. The data revealed that many of the problems identified were not specific to a single fidelity, for example in all three versions children had problems starting the game and re-starting the game. Therefore it is feasible to evaluate the usability of a game with children in the early stages of development using a low-fidelity prototype and have confidence in the results when transferring to a higher fidelity. However, care needs to be taken in the construction of the low-fidelity prototype as poor visuals may result in additional problems being identified that would not be found in the hi-fidelity versions.

7. Further Research

This study examined a single game for an iPad, in which the interaction could be simulated within a paper prototype. It would be interesting to see whether similar results are obtained if the level of interaction within the game is reduced. In the two paper versions, of the game used in this study, the children could physically turn over the cards to reveal the images. If the game required the use of touch or the accelerometer, and this was presented in the form of a storyboard, it is not known whether the children would still report positively.

There were limited usability problems reported and this may be attributed to the fact the game was reverse engineered from an already commercially available game. The study could be performed again with an early game idea, with sketches of a game that has not been fully built and a basic higher fidelity prototype of this game. It would be anticipated that additional usability problems might occur and depending on the interaction required, these might be fidelity related. For example issues that are pertinent to the device would not necessarily be identified if modeled on a paper prototype, such as the accelerometer.

8. ACKNOWLEDGMENTS

We thank the children who participated in the study from Hesketh-with-Becconsall All Saints CE Primary School. We would also like to thank the two PhD students Obelema Akobo and Chinedu Okwudili Obikwelu along with Matthew Horton who assisted with the data collection.

9. REFERENCES

[1] Shell, J. *The art of game design*. Morgan Kaufmann, 2008.

[2] Korhonen, H., Paavilainen, J. and Saarenpaa, H. Expert Review Method in Game Evaluations - Comparison of Two Playability Heuristics. In *Proceedings of the MindTrek 2009* (Tampere, 2009). ACM.

[3] Rudd, J., Stern, K. and Isesensee, S. Low vs high-fidelity prototyping debate. *Interactions*, 3, 1 (1996), 75-85.

[4] Buxton, B. *Sketching User Experiences - getting the design right and right design.* Morgan Kaufmann, San Francisco, 2007.

[5] Nielsen, J. and Molich, R. Heuristic evaluation of the user interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people* (Seattle, 1990). ACM.

[6] Desurvire, H., Caplan, M. and Toth, J. A. *Using heuristics to Evaluate the Playability of Games*. In *Proceedings of CHI* (Vienna, 2004). ACM.

[7] Pinelle, D., Wong, N., Stach, T. and Gutwin, C. *Usability Heuristics for Networked Multiplayer Games*. In *Proceedings of ACM 2009 International Conference on Supporting group work* (Sanibel Island, 2009). ACM.

[8] Korhonen, H. and Koivisto, E. M. Playability Heuristics for Mobile Games. In *Proceedings of the MobileHCI* (Helsinki, 2006). ACM.

[9] Sim, G., MacFarlane, S. and Horton, M. Evaluating Usability, Fun and Learning in Educational Software for Chiildren. In *Proceedings of EDMEDIA* (Montreal, 2005) AACE.

[10] Olsen, A., Smolentzov, L. and Strandvall, T. Comparing different eye tracking cues when using the retrospective

think aloud method in usability testing. In *Proceedings of the 24th British HCI Conference - Play is serious business* (Abertay, 2010). ACM.

[11] Liu, L. and Khooshabeh, P. Paper or interactive?: a study of prototyping techniques for ubiquitous computing environments. In *Proceedings of the CHI '03 Extended Abstracts on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA, 2003). ACM.

[12] Sauer, J. and Sonderegger, A. The influence of prototype fidelity and aesthetics of design in usability tests: Effects on user behaviour, subjective evaluation and emotion. *Applied Ergonomics*, 40 (2009), 670-677.

[13] Hanna, L., Risden, K. and Alexander, K. J. Guidelines for usability testing with children. *Interactions*, 4, 5 (1997), 9-14.

[14] Druin, A. *Cooperative inquiry: develping new technologies for children with children.* In *Proceedings of the CHI '03 Extended Abstracts on Human Factors in Computing Systems* (Pittsburgh, 1999). ACM.

[15] Markopoulos, P. and Bekker, M. On assessment of usability testing methods for children. *Interacting with Computers*, 15 (2003), 227-243.

[16] Reilly, D., Dearman, D., Welsman-Dinelle, M. and Inkpen, K. Evaluating Early Prototypes in Context: Trade-offs, Challenges, and Successes. *Pervasive Computing*, 4, 4 (2005), 42-50.

[17] Levi, M. D. and Conrad, F. G. A heuristic evaluation of a world wide web prototype. *Interactions*, 3, 4 (1996), 50-61.

[18] Lim, Y. K., Pangam, A., Periyasami, S. and Aneja, S. Comparative Analysis of High- and Low-fidelity Prototypes for More Valid Usability Evaluations of Mobile Devices. In *Proceedings of the NordiCHI* (Oslo, Norway, 2006). ACM.

[19] Snyder, C. *Paper Prototyping: The fast and Easy Way to Define and Refine User Interfaces*. Morgan Kaufmann, San Francisco, 2003.

[20] Kohler, B., Haladjian, J., Simeonova, B. and Ismailovic, D. Feedback in Low vs High Fidelity Visuals for Game Prototypes. In *Proceedings of the Games and Software Engineering* (Zurich, 2012). IEEE.

[21] Sefelin, R., Tscheligi, M. and Giller, V. Paper Prototyping - What is it good for? A comparison of Paper and Computer based Low fidelity prototyping. In *Proceedings of the Conference on Human Factors in Computing Systems* (Fort Lauderdale, 2003). ACM.

[22] Wiklund, M., Thurrot, C. and Dumas, J. Does the Fidelity of Software Prototypes Affect the Perception of Usability. In *Proceedings of the Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Atlanta, USA., 1992).

[23] Sauer, J., Franke, H. and Ruttinger, B. Designing interactive consumer products: utility of low-fidelity prototypes and effectiveness of enhanced control labeling. *Applied Ergonomics*, 39, 1 (2008), 71-85.

[24] Derboven, J., De Roeck, D., Verstraete, M., Geerts, D., Schneider-Barnes, J. and Luten, K. Comparing User Interaction with Low and High Fidelity Prototypes of Tabletop Surfaces. In *Proceedings of the NordiCHI* (Reykjavik, Iceland., 2010). ACM.

[25] Law, E. L.-C., Roto, V., Vermeeren, A. P. O. S., Kort, J. and Hassenzahl, M. Towards a Shared Definition of User Experience. In *Proceedings of the CHI 2008* (Florence, 2008). ACM Press.

[26] ISO *Ergonomics of human system interaction - Prt 210: Human-centred design for interactive system.* International Standards Organisation, 2010.

[27] Kaye, J. Evaluating experienced-focused HCI. In *Proceedings of the CHI 2007* (San Jose, 2007). ACM Press.

[28] Carroll, J. M. Beyond Fun. *Interactions*, 11, 5 (2004), 38-40.

[29] Inkpen, K. *Three Important Research Agendas for Educational Multimedia: Learning, Children and Gender*. In *Proceedings of the Educational MultiMedia* (Calgary, 1997).

[30] Malone, T. W. What makes things fun to learn? Heuristics for designing instructional computer games. In *Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems* (Palo Alto, 1980). ACM.

[31] Barendregt, W., Bekker, M. and Baauw, E. Development and evaluation of the problem identification picture cards method. *Cognition, Technology and Work*, 10, 2 (2008), 95-105.

[32] Read, J. C., MacFarlane, S. J. and Casey, C. *Endurability, Engagement and Expectations: Measuring Children's Fun*. Shaker Publishing, Amsterdam, 2002.

[33] Zaman, B. and Abeele, V. V. Laddering with Young Children in User Experence Evaluations: Theoretical Groundings and a Practical Case. In *Proceedings of the IDC* (Barcelona, 2010). ACM.

[34] Horton, M. and Read, J. C. Interactive Whiteboards in the Living Room? - Asking Children about their Technologies. In *Proceedings of the 22nd British HCI Conference* (Liverpool, 2008).

[35] Borgers, N., Leeuw, E. D. and Hox, J. Children as Respondents in Survey Research: Cognitive Development and Response Quality. *Buletin de Methodologie Sociologique*, 66 (2000), 60-75.

[36] Read, J. C. and Fine, K. Using Survey Methods for Design and Evaluation in Child Computer Interaction. In *Proceedings of the Interact 2005* (Rome, 2005).

[37] Sim, G. and Horton, M. Investigating children's opinions of games: Fun Toolkit vs This or That. In *Proceedings of the Interaction Design and Children* (Bremen, Germany., 2012). ACM.

[38] Hoysniemi, J., Hamalainen, P. and Turkki, L. Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interacting with Computers*, 15, 2 (2003), 203-225.

[39] Donker, A. and Reitsma, P. Usability Testing with Children. In *Proceedings of the Interacton Design and Children* (Maryland, 2004). ACM Press.

[40] Barendregt, W., Bekker, M., Bouwhuis, D. G. and Baauw, E. Predicting effectiveness of children participants in user testing based on personality charactersitics. *Behaviour & Information Technology*, 26, 2 (2007), 133-147.

[41] Polson, P., Lewis, C., Rieman, J. and Wharton, C. Cognitive walkthroughs: A method for theory-based evaluation of user interface. *International Journal of Man-Machine Studies*, 36 (1992), 741-773.

[42] Read, J. C. Validating the Fun Toolkit: an instrument for measuring children's opinion of technology. *Cognition, Technology and Work*, 10, 2 (2008), 119-128.