Investigating the Fidelity Effect when Evaluating Game Prototypes with Children

Gavin Sim
University of Central Lancashire
Preston, UK.
grsim@uclan.ac.uk

Brendan Cassidy
University of Central Lancashire
Preston, UK.
BCassidy1@uclan.ac.uk

The development and evaluation of prototypes is an important part of game development. Using an iPad, this study aimed to establish whether the fidelity of the prototype affects the ability of children to evaluate the user experience of a game. The participants were aged between 11 and 13 and used the Fun Toolkit to measure user experience in both fidelities. The results showed that the majority of children rated the low-fidelity version lower in terms of look, control and idea with the most significant difference being for the construct relating to the overall experience of the game. When evaluating monetary transactions with children it is important to realise that parental controls might influence the results.

Games. Fidelity Effect. User Experience. Children. Prototyping

1. INTRODUCTION

The game industry is a multi-billion dollar industry, with games being developed for a variety of platforms, devices and emerging technologies. Many of the games that are developed target children. It is therefore important to use an evaluation method that has been validated for use with children if user studies are performed. Adaptations to the evaluations are necessary as factors such as the decoration of the room, observational equipment and behaviour of the facilitator may affect the children's performance (Hanna et al., 1997).

Within the context of user experience, which has been defined by ISO, as a person's perception and responses that result from the use and/or anticipated use of a product, system or service (ISO, 2010), several evaluation methods have been developed for use with children including the Fun Toolkit (Read et al., 2002) and Problem Identification Picture Cards (Barendregt et al., 2008). These methods enable user testing of constructs such as fun to be evaluated within the context of games.

Companies face financial pressures to ensure rapid development of games and therefore it is critical that games get to market on time and are differentiated from their competitors. To aid the potential success of games, it is extremely important to playtest them as early and as often as possible during development. This is known to improve usability, and address game balancing and

motivation issues (Shell, 2008). Without early feedback the eventual player experience may not be optimal and players may switch to an alternative. If games are continually developed that are substandard the profitability of the company will be affected.

In the development stage gaming experience is usually evaluated once there is a prototype implemented that is ready for beta testing (Korhonen et al., 2009). Prototypes can exist in several forms – one categorisation of prototypes is as low or high fidelity. The development of low fidelity prototypes is usually associated with the use of material different from the final product for example, paper sketches (Rudd et al., 1996). High fidelity prototypes usually offer a level of functional interactivity using materials similar to those you would expect to find in a final product (Rudd et al., 1996). Time constraints and budgetary limitations often influence the fidelity of prototype being developed.

2. BACKGROUND AND RELATED WORK

There have been a number of comparative research studies of prototypes at different levels of fidelity. One study investigated the impact fidelity had on the users willingness to critique the interface as part of a usability study (Wiklund et al., 1992) and the results concluded that the number of usability problems is not affected by the prototype method. In contrast other studies highlight that whilst low fidelity (low visuals) prototypes can be

evaluated, the lack of refined graphics may bias evaluators against the products, despite the low fidelity being appropriate for usability testing (Kohler et al., 2012). However, another study concluded that users appeared to over compensate for deficiencies in aesthetics (Sauer and Sonderegger, 2009). There are studies showing results are equivalent to fully operational products and other studies reporting additional benefits of higher fidelity prototypes, therefore Yasar (2007) advocates the use of mixed fidelity prototypes which may be costly to develop.

It is noted from the literature that the majority of these studies have been performed with adults and it is unclear whether the findings can be generalized to children. In one study evaluating usability and user experience with children (Sim et al., 2013), a matching game of various levels of fidelity revealed no significant difference in user experience and the same usability problems were reported across the three prototypes. The limitations of this study were that it only analysed one game genre in which the children could physically interact with all prototype versions, and this is not always the case in game genres such as, first person shooters and platform games. Therefore there still remains uncertainty whether the results if obtained from a low fidelity prototype evaluated with children would infer to a higher fidelity due to the fidelity effect. This research aims to investigate the fidelity effect when evaluating games with children to understand the constructs that are consistent and different between prototypes.

3. METHOD

For this study, a within subject design methodology was adopted in which the user experience of a single game was evaluated in two different fidelities. The first version was a low fidelity prototype in the form of a storyboard and the other a functional game.

3.1 Participants

The participants were 27 school children from a UK high school, the children were aged 11-13 years old. The children took part in this study during a Mess Day at the author's institution (Horton et al., 2012). A Mess day consists of a group of school children coming into university to participate in a number of research activities. The first author of the paper acted as the facilitator during the study.

3.2 Game

For this study the Fix-it Felix game for the iPad3 was selected, as this game provided animations and interactivity that could not easily be simulated on paper, see Figure 1.

Figure 1: Functional Version of Game

To produce the lower fidelity version, the game was reverse engineered by capturing screen grabs of the actual game and tracing them using Adobe Illustrator, see Figure 2. Six screens were captured in this way to enable an accurate portrayal of the game and this approach has been used successfully in another study (Sim et al., 2013). Text was then added to make a storyboard to explain the game concept and to outline the interactions required to play the game. The annotated screens were then saved as a PDF to be displayed in a linear way on the iPad.

You control the character with the joystick in the bottom left corner. The character must jump from window to window fixing the broken glass. To fix the glass you press the hammer button on the right which fixes the window.

The man at the top throws bricks at you from above. If the brick hits you you lose a life and you have 3 lives in total.

There is a time limit to fix all the windows in.

On the first level you have to fix all the windows on the first 3 floors.

Figure 2: Low-Fidelity Version of Game

Despite the fact that low-fidelity prototypes are normally 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 and this is important to reduce confounds.

3.3 Study Design

The study aimed to establish whether the fidelity of the prototype affected the ability of children to evaluate user experience of a game designed for the iPad.

To measure user experience the decision was made to use an adaptation of the Fun Toolkit (Read et al., 2002). The first tool, the Smileyometer, is a visual analogue scale with a coding based upon a 5 point scale, see Figure 3.

Figure 3: Smileyometer rating scale

The Smileyometer is usually used before and after 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 their experience. In total five questions (1 before and 4 after interaction) were asked that were intended to be answered by using the Smileyometer scale; these were:

Before:

- Q1 - What do you think the game will be like?

After:

- Q2 - The idea of the game (fixing windows) is:
- Q3 - The way you control the character and fix windows is:
- Q4 - How do you think the game will look (image the screens made with colour):
- Q5 - Overall I think the game is:

For the iPad version, the wording of the question relating to look was altered to simply state *I think the game looks*, as the children were viewing the completed graphics. For the first Smileyometer question they were asked to comment why they had selected that particular option.

In addition to these questions, the Again Again table was adapted. The Again Again table is a table that requires the children to pick 'yes', 'maybe' or 'no' for each activity they have experienced. In this case the children were asked:

- If the game were free would you download it from the app store?
- If the game were 69p would you download it from the app store?

For these two questions they had to select yes, maybe or no. They were also asked to comment why they selected this option.

3.4 Apparatus

The children used an iPad3 to view and interact with the prototypes. For the Fun Toolkit the researcher gave the children a pen and a data capture form to complete the Smileyometers and Again Again tables.

3.5 Procedure

The research was conducted in a computer laboratory within the university. The children came in groups of 4 and were allocated to a desk containing the questionnaires and iPad. The order the children interacted with the prototype and game were counterbalanced, with half the children playing the game first followed by the low-fidelity version and vice versa. Depending on the order either the storyboard version of the game was on display when the child turned on the iPad or the first screen of the game. Before the child viewed the entire storyboard or played the game, four initial questions were asked about their age, experience of using the iPad and whether they had played this game. They then answered the first question of the Smileyometer regarding their expectations after they had seen the first screen of the game.

The children would then go through the storyboard reading the information on the screens and once complete, they were required to answer the remaining questions. They would follow the same procedure for the high fidelity version of the game until they reached level 4 (for consistency with the low fidelity version) when they were asked to complete the survey again. The whole procedure lasted about 20 minutes per group.

3.6 Analysis

All the children managed to complete all the questions in the survey tool. The questions using the Smileyometer, see Figure 3, were coded in an ordinal way 1- 5, where 5 represented Brilliant and 1 Awful. For the Again Again table, yes was coded as 2, maybe as 1 and no as 0.

In line with other studies using this scale arithmetic averages have been taken of these scores in order to show differences (Read, 2012). A full analysis of the qualitative data is beyond the scope of this paper, however, some of the comments from the children are used to support the discussion.

4. RESULTS

Of the 27 children, 19 had used an iPad before and 2 children had played the game before.

The results for the first question of the Smileyometer, before the children had interacted

with either the game or prototype is presented in table 1 below:

Table 1: Mean scores and standard deviation for the first question relating to expectations

Question	Low-Fidelity		Game	
	Mean	SD	Mean	SD
Q1	3.22	.974	3.52	1.01

A t-test revealed no significant difference between the Smileyometer results for Q1, $t=-1.551$, $df=26$, $p=.133$.

The results of the four questions after they had interacted with the prototype or game are presented in table 2.

Table 2: Mean scores and standard deviation for the four question following interaction

Question	Low-Fidelity		Game	
	Mean	SD	Mean	SD
Q2	3.15	.77	3.67	1.01
Q3	3.30	.87	3.85	1.21
Q4	3.56	1.01	3.89	.93
Q5	3.59	.93	4.11	1.05

A t-test was performed for the four questions following the interaction and in this case the results showed that there were three weakly (shown below) and one strongly significant difference. For the question relating to the game idea the results from the t-test were $t=-3.017$, $df=26$, $p=.006$, for control $t=-2.658$, $df=26$, $p=.013$, regarding the look $t=-2.55$, $df=26$, $p=.017$ and finally the overall rating of the game $t=-3.578$, $df=26$, $p=.001$.

In response to the questions as to whether the child would download the game if it were free or 69p, the results are presented in table 3. The number in brackets represents the score from the low-fidelity prototype for the question, whilst the other number is for the actual game.

Table 3: Frequency of responses to whether the child would download the game if it was free or 69p

	Yes	Maybe	No
Free	(17) 23	(9) 2	(1) 2
69p	(6) 9	(11) 10	(10) 8

5. DISCUSSION

The children's initial expectations of the game showed no difference between the actual game and the low-fidelity prototype. Analysis of the children's comments suggest that they associated the game with the Disney film *Wreck it Ralph*, which was showing in cinemas at the time of the

study, and so may have perceived the game to be of good quality, other comments related to the fact that it looked fun and interesting.

For the questions after they had played the game there were clear differences. These differences may have been caused by the quality and detail within the storyboard. The questions relating to the game idea and control had the lowest mean scores in the low-fidelity prototype. One possible cause for this may be that it might have been difficult for the children to clearly understand the game idea and how they would interact with game from a static prototype.

Regarding aesthetics the low fidelity prototype was lower than the actual game and this might have been expected. The low fidelity prototype was purely in black and white. Some children may have construed the lack of colour as a negative aspect, with comments like:

- It is not colour on it
- It looks a bit basic
- The game doesn't look good

Sauer and Sonderegger (2009) suggested that user emotions are more positively affected by the attractiveness of the technology. Given the fact that both versions of the game were delivered on the iPad it is unlikely that this would have been a cause of the difference.

The largest difference was found in the overall rating of the game, with children on average rating the iPad game as really good. The difference may have been attributed to the fact they could not physically play the game on the low-fidelity version. In a study of a matching game the children could physically play the low-fidelity version and this resulted in there being no significant difference between the versions after they had played the game (Sim et al., 2013). However it may simply be down to the fact that the children struggled to imagine how the game mechanics would work, as well as finding it difficult to visualise the game play.

To further analyse these differences, the original data was further examined to determine the size of change for each of the questions asked. Table 4 below shows the difference in rating between the iPad game and the low-fidelity prototype. In the table, a positive score of 1 indicates that the child rated the iPad game higher than the low-fidelity version by 1 rating on the Smileyometer. For example, the low-fidelity prototype might have been rated *Good* but the actual game was rated *Really Good*. Whilst a negative score indicates that the low fidelity prototype scored higher than the actual iPad game for that particular question.

Table 4: Frequency of change for each question between the iPad game and Low Fidelity Prototype

	Children's difference between fidelity versions					
	-2	-1	0	1	2	3
Q1	0	5	13	6	2	1
Q2	1	1	11	11	3	0
Q3	1	3	8	11	3	1
Q4	0	3	12	12	0	0
Q5	0	1	14	9	3	0

For questions 2-5, approximately 50% of the children rated the game higher by at least 1 rating than the low fidelity version. This result suggests that a large portion of children underestimated the game based on their interaction with the low-fidelity prototype.

With regards to analysing the responses to whether the children would download the game if it were free or 69p, there is clearly a large change in responses between the low-fidelity prototype and the game. For the iPad game, out of 27 children, 24 stated that they would download the game if it were free, however only 9 stated they would be prepared to pay for the App. In contrast the results were lower for the low-fidelity version with only 17 of the children stating they would be willing to download the game if it was free and 6 if it was 69p. The children were also asked to comment on why they selected this response and typical comments from the free and paid for versions after they had interacted with the game are reproduced below:

iPad Game Free

- It's a fun game to play, slightly challenging
- It was fun to play
- I download most free apps and delete them if they are bad

iPad Game Paid

- As good as it is, I don't think it's good enough for 69p. Making the graphics better would improve it.
- I am not allowed to spend money in the app store.
- I would like to try a free version first

Some of the children reasons for their choice based on viewing the low-fidelity are reported below:

Low-Fidelity Free

- So that I can see what it will be like to play this game
- It will be good to play
- Because it could get boring after a while

Low-Fidelity Paid

- I may not be allowed to get it and it may be difficult
- I don't get things that you have to pay
- Because it is not worth to buy, it's worth to get it for free

Some of the reasons for selecting the option *no* for whether they would download it if it were 69p came from parental restrictions on the use of app stores. If they were allowed to download games then this response might have been different. It is evident that the majority of children would be willing to download the game if it was free after being exposed to the low-fidelity prototype or actually playing the game.

6. CONCLUSIONS

This paper aimed to investigate the constructs that are consistent and different between prototypes when evaluating games with children. In this study there was no difference in children's initial expectation of the game before they interacted with the prototype or actual game. Despite this, there were significant differences for the other constructs investigated. It may be that for games developers it is possible to evaluate user experience with children with a caveat that the results may be lower than the actual game if a high-fidelity or functional game was evaluated.

When evaluating monetary transactions with children it is important to realise that parental controls might influence the results. The children might be reporting the answer based on parental influence rather than their actual feelings.

The results might have been influenced by the quality of the low-fidelity prototype and the information provided.

7. FURTHER RESEARCH

This study examined the fidelity effect of a single game using one style of prototype, in this case a storyboard. It would be interesting to see whether similar results are obtained for the same game with a different style of prototype or a different game genre is evaluated. For example, the text describing the game might be adapted or the prototype might be presented on paper rather than on the iPad.

It could be worthwhile influencing some of the constructs analysed within this study such as, the aesthetics through the inclusion of colour and

seeing if this affected the responses for question 4 in the survey.

To investigate the change in children's responses between the two versions of the game, it might be useful to analyse a game that is judged to be poor or not very fun. This would give an insight into frequency change and whether the low-fidelity prototype still scored lower than the higher-fidelity version.

In this study data relating to gender was not collected and therefore it might be worthwhile determining whether gender influences the results.

8. REFERENCES

- Barendregt, W., Bekker, M. & Baauw, E. (2008). Development and evaluation of the problem identification picture cards method. *Cognition, Technology and Work*, 10, (95-105).
- Hanna, L., Risdén, K. & Alexander, K. J. (1997). Guidelines for usability testing with children. *Interactions*, 4, (9-14).
- Horton, M., Read, J. C., Mazzone, E., Sim, G. & Fitton, D. (2012). School friendly participatory research activities with children. *CHI 12 Extend Abstracts*. Austin Texas: ACM.
- Iso (2010). Ergonomics of human system interaction - Part 210: Human-centred design for interactive system. Switzerland: International Standards Organisation.
- Kohler, B., Haladjian, J., Simeonova, B. & Ismailovic, D. (2012). Feedback in Low vs High Fidelity Visuals for Game Prototypes. *Games and Software Engineering*. Zurich: IEEE.
- Korhonen, H., Paavilainen, J. & Saarenpää, H. (2009). Expert Review Method in Game Evaluations - Comparison of Two Playability Heuristics. *MindTrek 2009*. Tampere: ACM.
- Read, J., Macfarlane, S. & Casey, C. (2002). Endurability, Engagement and Expectations: Measuring Children's Fun. *Interaction Design and Children*. Eindhoven.
- Read, J. C. (2012). Evaluating artefacts with children: age and technology effects in the reporting of expected and experienced fun. *14th ACM International Conference on Multimodal Interaction*. Santa Monica: ACM.
- Rudd, J., Stern, K. & Isesensee, S. (1996). Low vs high-fidelity prototyping debate. *Interactions*, 3, (75-85).
- Sauer, J. & Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests: Effects on user behaviour, subjective evaluation and emotion. *Applied Ergonomics*, 40, (670-677).
- Shell, J. (2008). *The art of game design*, Morgan Kaufmann.
- Sim, G., Cassidy, B. & Read, J. C. (2013). Understanding the Fidelity Effect when Evaluating Games with children. *Interaction Design and Children*. New York: ACM.
- Wiklund, M., Thurrot, C. & Dumas, J. (1992). Does the Fidelity of Software Prototypes Affect the Perception of Usability. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Atlanta, USA.
- Yasar, A.-U.-H. (2007). Enhancing Experience Prototyping by the help of Mixed-Fidelity Prototypes. *Mobility*. Singapore: ACM.