# Feedback in low vs. high fidelity visuals for game prototypes

Barbara Köhler, Juan Haladjian, Blagina Simeonova, Damir Ismailović, Bernd Brügge
*Institut für Informatik, Technische Universität München*
*Garching bei München, Germany*
*Email: {koehlerb, haladjian, simeonob, damir.ismailovic, bruegge}@cs.tum.edu*

***Abstract*—Prototypes have proven to be a good practice in different areas. In the gaming industry, they help identify usability and gameplay issues, among others. The earlier these issues are identified, the less effort is required fixing them. But game assets like graphics are often expensive and are available later on, or even after game functionality has been implemented. Game prototypes are in this case created using lower fidelity visuals. While this technique makes it possible to perform usability tests, it may bias the feedback provided by usability testers. In this paper we investigate how the fidelity of the prototypes used for usability testing influence the feedback provided by testers.**

***Keywords*-game engineering; serious games; style; styling;**

## I. Introduction

Prototyping is an important technique to elicit information about a design and determine its usability and feasibility as early in the development process as possible. However it is important to select the correct prototypes for the momentary step of development. Otherwise one might receive less information or different information from what was expected. It has been shown in several studies that the fidelity of prototypes for software can heavily influence the amount of work required to create it and the kind of feedback that one receives.

[1] describe several reasons why researchers think using low-fidelity prototypes are a useful tool in identifying usability issues. It is claimed that they are (1) an efficient way to search the design space [2], (2) predictive of preferences in the actual product [3], (3) enhance user participation in the design process [4], enable visualization of possible design solutions [5] and provoke innovation [6]. The usefulness of low-fidelity prototypes can also be seen in successful usability studies conducted with paper prototypes [7]. However it is important to note that low- and high-fidelity prototypes are still very broad categories, which can still differ by a lot. In software the fidelity of a prototype can independently varied along four axes [1]:

- Depth: extent to which details of its operation are complete
- Breadth: number of features the prototype supports
- Similarity of interaction: how one communicates with the product (pressing buttons, tapping the screen, etc.)
- Visual representation: aspects of the product that do not directly influence its functionality, such as choice of colors and graphic design

Reducing fidelity on each of those axes can successfully reduce development cost. However at the same time the validity of results from a usability study with those prototypes might be influenced. For games the biggest cost factor lies often in creating all the assets required. Therefore it is often crucial to build prototypes with assets of a significantly lower quality than the one that will be present in the final game. However especially for games it is extremely important to playtest them as early and as often as possible in order to improve not only on their usability but also to address game balancing and motivation issues as early as possible [8].

For utility applications there is already some research that indicates that a reduction of visual fidelity will not reduce the number of usability issues found, some studies and experiences of developers even indicate that there will be less low-level feedback about the visual aspects of the design which is only very useful in the final stages of the development, when one specifically wants to receive feedback on the visual design. The effect that testers focus on visuals if they look very polished has also been reported by practitioners, e.g. in [7]. "When something appears to be finished, minor flaws stand out an will catch the user's attention. To put it another way, people will nitpick." However we could not find similar studies that justified a similar approach for games. During the development of our games we were forced to develop with low-fidelity graphics. We used them for user testing from the first stages in order to continually improve our design. Now that the development of the games is close to being finished we want to evaluate the validity of our approach.

The purpose of this paper is twofold. First we present a case study that originated the topic of this paper. Then we describe and present an evaluation methodology and preliminary results for measuring how the fidelity of a game prototype influences the user feedback.

## II. Case Study

We developed an industrial game project in a period of two years of time. The game targeted kids in the age of four to seven years. The purpose of the game is teaching basic mathematics; counting, drawing digits, recognizing numbers, basic addition and subtraction, etc. The author of the game

series originated the game idea. A team of two professional developers, one team leader, and eight student developers developed the game. The game targets the iPad platform.

During the first meeting between the development team and the clients, the author of the game series presented a series of hand drawn sketches and explained the entire game ideas. These sketches were then photographed and different game actors were cut off and included in game prototypes. Missing images were quickly drawn by hand, scanned and included in the game. We then used the prototypes to perform usability tests with a wide variety of users; some of them were kids, some others were software engineers. We also held discussions with experts in the field, including the author of the game, who is a pedagogue. A few months afterwards, when the polished images were finished, we integrated them and repeated the usability tests. This process is illustrated by Figure 1.

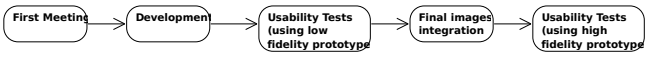


Figure 1. Development process we followed.

During the usability tests, we were able to find not only usability problems, but also gameplay problems. Some kids did not pay atention to game explanations and did not know what to do, instead they tried randomly, eventually solving the task. But by doing so, they completely avoided learning what the game was actually meant to teach. We also observed that some kids found some tasks too long and were just lazy to continue playing. One of the testers had difficulties using the iPad's accelerometer, she preferred moving it always in one axis, which caused her not to explore entire parts of the game.

We noticed a lot of time was invested in usability tests. On one hand, it was hard to organize meetings with nursery schools and wait for authorization from parents. On the other, some parts of the game were hard to understand by kids, who needed a long time to grasp it without our help.

Since usability tests were costly, we wanted to maximize the profit out of them. For example, we were not sure whether the usability problems identified using the low fidelity prototype were valid at all, as they could have been influenced by the low quality of the graphics. When the kids were for example bored and lazy to continue playing, up to what extent were low quality assets the cause? We also did not know what to expect from the usability tests in each case. Do users testing a game with non-polished graphics criticize it more because it does not meet their expectations? Or they criticize less because they assume it is still under development? How severe are the issues found in each case? These questions originated the study presented in this paper.

## III. Related Work

The question how visual fidelity of a prototype influences results of usability tests has already been dealt with in a series of studies. Most of them try to measure a variety of effects beginning with the relation between perceived visual appeal and perceived usability. They also measure the type and amount of user feedback received, e.g. a focus on higher level feedback in low-fidelity prototypes, and the number of usability errors detected.

The perceived attractiveness of a prototype can significantly influence the perceived usability reported by the users. Usually a more visually pleasing prototype is automatically perceived to have a higher ease of use. This effect has been observed in prototypes of websites ( [9], [10]) and computer software ( [11]). In those studies, however, the aesthetics were not manipulated experimentally but two different products aiming at the same goal were used. As a result, aesthetics and usability might have been confounded. Further studies have tried to prove this effect by manipulating the aesthetics experimentally and by avoiding differences in the prototypes concerning other dimensions, e.g. color settings of a webpage ( [12]), the manipulation of the shape of an electronic phonebook-simulator ( [13]), the variation in the design of a webpage (Brady and Phillips, 2003), and the manipulation of the color of a mobile phone's case and screen ( [14], [15]). One possible explanation for the connection between perceived aesthetics and perceived usability can be seen in the halo-effect. The halo-effect is a cognitive bias where specific, salient characteristics of a specific object influence the perception of other characteristics. Edward Thorndike has first supported this effect with empirical research.

The influence of aesthetics does not stop at perceived usability, measures of user performance like the number usability errors made, task completion time and interaction efficiency (ratio between number of user inputs and optimal number of user inputs) are also discussed. While some studies measured a positive outcome of a visually pleasant interface on those performance measures [15] others even detected decrements in performance ( [13], [14]). Finally, some could not detect any measurable effect ( [9], [16]). When looking for a theoretical explanation for both orthogonal effects there seem to be two opposing forces. On the one side the nice appearance increases the motivation to work with the software ( [17]). On the other, the user interface can make the experience so much fun that the user starts toying around with the system, spending more time with a task than is actually required.

Additional to objective measures the feedback given by users about a prototype is also extremely important. But the type of feedback might also vary with the fidelity of the graphics used. A study in Wikilund[9] indicates that the fidelity did not affect the sensitivity of the usability tests, but

that the kinds of problems one can detect are very different. Some studies ( [18], [19], [20]) report a focus on higher-level issues in low-fidelity prototypes.

## IV. Experiment Design and Preliminary Results

### A. Hypothesis

Looking at the large body of research we built a number of different hypothesis for the influence of the visual fidelity on games.

- $H1$ : Low-fidelity prototypes provoke more feedback on functionality
- $H2$ : High-fidelity prototypes lead to the user focusing on minor details in the prototype
- $H3$ : A prototype with a graphics look finished will be perceived as significantly more finished overall

We assume that low-fidelity prototypes encourage players to give more feedback on the functionality of the game while high-fidelity prototypes will mainly provoke feedback on minor details. This could be due to the fact that in the low-fidelity prototypes there are less signals that divert the player from the actual prototype. Furthermore if the visuals look unfinished the players assume that they will be improved anyways and that they are not supposed to critique them right now. Last but not least the unfinished look might be a signal that not only the graphics but also everything else is till not completely worked out. Therefore it seems more socially acceptable for the user to criticize the prototype.

- $H4$ : The amount of usability issues which can be detected with high- and low-fidelity prototypes do not vary significantly

As it is expensive to develop the final assets for a game they are usually only developed as soon as the game is nearly finished. However at this point it is too late to change anything essential in the game mechanics and the layout. Therefore problems in the game need to be detected before the graphics are completed. We want to show that many issues actually can be found even when testing with unfinished graphics and that it will not reduce the amount of usability issues encountered.

- $H5$ : Perceived usability is higher when perceived visual attractiveness is higher

As stated earlier the visual attractiveness of a prototype can influence the overall perception of a prototype. We think that this effect is also present in prototypes for games. It might be even stronger than in business software as the graphics present a much more significant part of the game in comparison to other software, where functionality and efficiency are more central as the main goal of the software is to work towards reaching a goal and maybe creating a product.

- $H6$ : Usability testing with adults will reveal the same issues as testing with children

We found that testing with children comes at a much higher cost and effort than testing with children. While balancing game difficulty should only be done with the actual target group we believe that many usability issues are essentially the same when testing with adults. Therefore doing pretests with adults to remove the most cumbersome usability issues will our tests with children more effective.

### B. Method

*1) Participants:* Six students, aged between 20 and 28, participated in the study. Two of the participants were female, four male. They all had very different levels of prior experience with touch devices. While some had a touch device, e.g. iPhone, Android etc., that they used daily the other three students did not own any such device. They also did not have any other experience with mobile devices. The study was mainly conducted with adults although the game is clearly meant to be played by children. However as mentioned earlier studying children raised some problems not present when testing adults.

*2) Experimental Design:* The experiment was done as a 2x2 mixed design with aesthetics of design as a between-subject variable. Participants were randomly assigned to the group playing with the high-fidelity or the low-fidelity prototype first. They were then asked to play with the prototype while trying to reach certain goals. During the whole process the participants were asked to think aloud what they are trying to do and how they think they can achieve it. An observer made notes about this as well as other observations made during the experiment. Additionally the experiments were recorded if the researchers wanted to look at one of the sessions again at a later point in the study. After the player had completed one guided play through he was asked a set of questions about the perceived usability and attractiveness of the design. After that a second play through with the other prototype was done in order to see whether participants changed their behavior when confronted with a different design for the same game.

*3) Measures and instruments:* In the study all participants were observed while playing the game. Everything interesting behavior that they displayed as well as comments or questions was noted. Two different researchers then used those protocols to compile a list of usability errors that could be deduced from the observations. Each researcher furthermore sorted the issues by severity according to the scale suggested by [21].

0 = I don't agree that this is a usability problem at all

1 = Cosmetic problem only: need not be fixed unless extra time is available on project

2 = Minor usability problem: fixing this should be given low priority

3 = Major usability problem: important to fix, so should be given high priority

4 = Usability catastrophe: imperative to fix this before product can be released

In addition to those observations the number of times when the players make a request for help by tapping Emil and Pauline was also counted. In comparison to some of the papers above we do not count the number of wrong taps however, as in games the goal is to have fun. So while an additional tap on one of the items in the labyrinth for example will not lead to a direct solution of the problem posed by the game it nevertheless is a valuable interaction with the system. Therefore it is very hard to draw the line as to whether an action was necessary and correct when talking about games.

In addition to the data collected by the researchers we also designed a small questionnaire that asks for personal data like age, current occupation and prior experience with touch devices and games. Furthermore the user fills out another small survey for each prototype presented to him, rating perceived usability, perceived attractiveness, overall appeal of the game and estimated completion in %.The questionnaire uses a 5 point Likert scale.

*4) Materials:* The game used in the study is part of a game series called Emil and Pauline published by USM. It is targeted towards children of age 4 to 7. In the first screen of the game the player is supposed to log in by drawing a secret sign that he can later use to load his game state(Figure 2). At the top of the cave there are the already existing accounts that the player can choose from. In every screen tapping the two main characters Emil and Pauline will result in Emil and Pauline explaining the current task at hand. After the login screen the player enters a cave, where he can select between different games that teach different topics in the area of mathematics. For the study two of the five minigames were selected. In the first one the player needs to navigate a labyrinth steering the character with the accelerometer and collecting numbers. Add the end of the labyrinth the player needs to select the sum of all the numbers he collected out of several other numbers. In the second game the player needs to distinguish geometrical shapes by three different properties: shape (square, circle, triangle), Color (red, green, blue), and size (small or big).

Two prototypes of the game were created, that are equal but for the graphics used. The graphics used in the low-fidelity variant of the prototype were similar to the finalized graphics, but reduced in resolution without colors (unless the colors were important to understand and play the game) and only consisting of contours. In some scenes colors were added, when they were required in order to play the game. The layout for each game scene was kept identical so that both prototypes could be played in the exactly same way.

*5) User tasks:* During evaluation the users were asked to solve a set of tasks. The first one was to create a new account and log in using it. After that the user was supposed to start one of the mini games. As soon as they had one the first

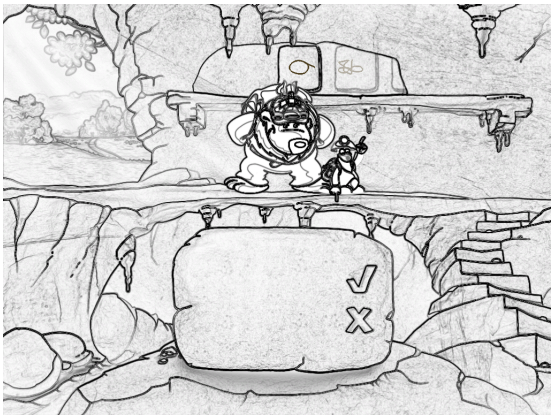

Figure 2. Low-Fidelity prototype

Figure 3. High-Fidelity prototype

mini game they had to start a second minigame, which they played until they won. Finally they were asked to restart the game and remove their newly created account. The same set of tasks was done with the second prototype.

*6) Procedure:* The study was conducted at the desk of a private room. Participants were volunteers from a university dorm and from the computer science building. Participants were randomly assigned to one of the experimental conditions. Participation in the study took between 20 to 40 minutes (mainly depending on prior experience with touch devices and games).

At the beginning of the test the participants were welcomed by the experimenter and informed that they would take part in a usability test for a learning game currently in development. They then filled out a short questionnaire asking for their age, prior experience with touch devices in general and games in particular. After that they then played the same game a second time with the other graphics. We did this as we wanted to see whether a person changes their mind on seeing the other prototype and also because we assumed that some persons would more strict or more lenient with their criticism independently of prototype style. Testing

every person in both conditions could help us to measure this effect.

### C. Results

The number of participants that we managed to raise for an initial test of our study design is most certainly to small to be able to definitely prove something. However the data already indicates some things. The perceived graphic attractiveness was in general higher in the high-fidelity prototype (average in low-fidelity prototype: 3 in high-fidelity prototype: 4.5). This was to some extent a smaller difference than we expected, but it when we asked the participants they usually stated that the layout and character design already looked relatively polished and consistent.

Also we were able to find a magnitude of usability issues in both conditions. Mostly it did not seem to matter which graphics were used. The problems the users struggled with were mostly the same.

When considering our hypothesis that we will receive more feedback targeted towards details only in the high-fidelity prototype however we did not make any supporting observations. We think that this can be due to the fact that players do not have a fixed expectation of what a user interface needs to look like. In games the graphics are usually much more individual than in other software, where there are exact expectations for the locations and alignment of specific UI Elements. Accordingly the testers do not find so many details that they can and want to criticize.

Finally the usability issues found correspond closesly to some of the largest problems that we had when testing our prototype with children. This indicates that indeed for initial usability testing, testing with adults even for a child's game might be acceptable.

### D. Validity and Limitations

In this study we only tested 3 persons in each condition. While this might seem very little at first glance it is considered sufficient regarding the number of usability issues one can find. Tom Landauer and Jakob Nielson showed that the number of usability problems discovered in a test with n users is $N(1-(1-L)^n)$ where L is the percentage of problems found by testing with one user which is typically around 30% [22]. Considering this information one can calculate that about 85% of the issues can be found with 5 testers only, with three users we still are able to detect about 65%. After more than 5 users they claim that one will mostly observe a repetition of the problems already seen with the other users. While the low number of participants in the study can be justified considering the comparison of amount of usability issues detected, the findings considering the type of feedback received might be harder to defend. In our study we did not find any conclusive results considering the type of feedback. We neither observed a focus on visual details in prototypes with high visual refinement nor did we observe a bigger willingness to criticize our prototype that looked less finished. It might be possible that these effects can only be measured by observing a larger amount of participants.

## V. Conclusion

So far we designed a study and already collected some first results to evaluate how the aesthetics of the graphics influence the feedback. Also we targeted the question whether children and adults find similar problems when using an interface. We think that the amount of usability issues that are found does not entirely rely on using the intended audience, at least for a serious game targeted towards children. Severe problems can also be found by testing with adults which can significantly speed up initial usability tests as less preparation is required in comparison to testing with children. Furthermore most usability issues can be found no matter what kind of visual refinement is used. Therefore especially in the beginning of a project one should only use placeholder graphics and start testing with them. This increases the number of iterations that can be done overall improving the game while at the same time keeping costs for graphic creation comparatively low. We cannot support the thesis that users focus on minor visual details. The participants of our usability tests did almost never refer to the visuals, neither in the low-fidelity nor in the high-fidelity condition. This might be due to the fact that users expect an individual style in a game and therefore do not have a detailed expectation of the look the game is supposed to have. Another explanation could be that the questions used in the usability tests strongly guided the participants to focus on mechanics instead of visuals.

## References


[1] R. a. Virzi, J. L. Sokolov, and D. Karis, "Usability problem identification using both low- and high-fidelity prototypes," *Proceedings of the SIGCHI conference on Human factors in computing systems common ground - CHI '96*, pp. 236–243, 1996. [Online]. Available: http://portal.acm.org/citation.cfm?doid=238386.238516

[2] R. A. Virzi, "What can you learn from a low-fidelity prototype?" in *Proc. of Human Factors Society 33rd Annual Meeting*, 1989, pp. 224–228.

[3] M. Wiklund, C. Thurrott, and J. Dumas, "Does the fidelity of software prototypes affect the perception of usability?" *Proceedings of the Human Factors Society 36th Annual Meeting*, pp. 399 – 403, 1992. [Online]. Available: http://www.citeulike.org/user/mvolger/article/2939282

[4] M. J. Muller, "An exploration in participatory design," in *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91*. New York, New York, USA: ACM Press, Mar. 1991, pp. 225–231. [Online]. Available: http://dl.acm.org/citation.cfm?id=108844.108896

[5] B. Moggridge, "Design by story-telling," *Applied Ergonomics*, vol. 24, no. 1, pp. 15–18, Feb. 1993. [Online]. Available: http://dx.doi.org/10.1016/0003-6870(93)90154-2

[6] W. Wulff, S. Evenson, and J. Rheinfrank, "Animating interfaces," in *Proceedings of the 1990 ACM conference on Computer-supported cooperative work - CSCW '90*. New York, New York, USA: ACM Press, Sep. 1990, pp. 241–254. [Online]. Available: http://dl.acm.org/citation.cfm?id=99332.99358

[7] C. Snyder, *Paper Prototyping: the fast and easy way to design and refine user interfaces.* San Francisco, CA: Morgan Kaufmann, 2003.

[8] J. Schell, *The Art of Game Design*. Elsevier, 2008.

[9] J. Hartmann, A. Sutcliffe, and A. D. Angeli, "Investigating Attractiveness in Web User Interfaces," *Design*, pp. 387–396, 2007.

[10] B. N. Schenkman, B. N. Schenkman, and F. U. J. Nsson, "Aesthetics and preferences of web pages," *BEHAVIOUR & INFORMATION TECHNOLOGY*, vol. 19, pp. 367 – 377, 2000. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.130.1138

[11] M. Hassenzahl, "The Interplay of Beauty , Goodness , and Usability in Interactive Products," vol. 19, pp. 319–349, 2004.

[12] I. Nakarada-Kordic and B. Lobb, "Effect of perceived attractiveness of web interface design on visual search of web sites," in *Proceedings of the 6th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction making CHI natural - CHINZ '05*. New York, New York, USA: ACM Press, Jul. 2005, pp. 25–27. [Online]. Available: http://dl.acm.org/citation.cfm?id=1073943.1073949

[13] T. Ben-Bassat, J. Meyer, and N. Tractinsky, "Economic and subjective measures of the perceived value of aesthetics and usability," *ACM Transactions on Computer-Human Interaction*, vol. 13, no. 2, pp. 210–234, Jun. 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1165734.1165737

[14] J. Sauer and A. Sonderegger, "The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion." *Applied ergonomics*, vol. 40, no. 4, pp. 670–7, Jul. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18691696

[15] A. Sonderegger and J. Sauer, "The influence of design aesthetics in usability testing: effects on user performance and perceived usability." *Applied ergonomics*, vol. 41, no. 3, pp. 403–10, May 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19892317

[16] M. Thuring and S. Mahlke, "Usability, aesthetics and emotions in human-technology interaction," *International Journal of Psychology*, vol. 42, no. 4, pp. 253–264, Aug. 2007. [Online]. Available: http://dx.doi.org/10.1080/00207590701396674

[17] G. Lindgaard, "Scientific Commons: Aesthetics, visual appeal, usability and user satisfaction: what do the user's eyes tell the user's brain?" Aug. 2007. [Online]. Available: http://en.scientificcommons.org/51861015

[18] A. BLACK, "Visible planning on paper and on screen: The impact of working medium on decision-making by novice graphic designers," *Behaviour & Information Technology*, vol. 9, no. 4, pp. 283–296, Jul. 1990. [Online]. Available: http://dx.doi.org/10.1080/01449299008924244

[19] J. Landay and B. Myers, "Sketching interfaces: toward more human interface design," *Computer*, vol. 34, no. 3, pp. 56–64, Mar. 2001. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=910894

[20] Y. Y. Wong, "Rough and ready prototypes," in *Posters and short talks of the 1992 SIGCHI conference on Human factors in computing systems - CHI '92*. New York, New York, USA: ACM Press, May 1992, p. 83. [Online]. Available: http://dl.acm.org/citation.cfm?id=1125021.1125094

[21] J. Nielsen, "Severity Ratings for Usability Problems," *Internet URL httpwww useit compapersheuristicseverityrating html Version*, vol. 2008, pp. 11–99, 1995. [Online]. Available: http://www.useit.com/papers/heuristic/severityrating.html

[22] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93*. New York, New York, USA: ACM Press, May 1993, pp. 206–213. [Online]. Available: http://dl.acm.org/citation.cfm?id=169059.169166