Evaluating Game Preference using the Fun Toolkit across Cultures

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

Matthew Horton
University of Central Lancashire
Preston
mplhorton@uclan.ac.uk

Nicky Danino
University of Central Lancashire
Preston
NDanino@uclan.ac.uk

Over the past decade many new evaluation methods have emerged for evaluating user experience with children, but the results of these studies have tended to be reported in isolation and cultural implications have been largely ignored. This paper reports on a comparative analysis of the Fun Toolkit and the effect of culture on game preference. In total 37 children aged between 7 and 9 participated in the study, from a school in the UK and Jordan. The children played 2 different games on a tablet PC and their experiences of each were captured using the Fun Toolkit. The results showed that culture did not appear to affect children's preference and Fun Toolkit is a valid user experience tool across cultures.

User Experience, Evaluation Methods, Child Computer Interaction, Culture.

1. INTRODUCTION

Evaluation methods play a vital role in HCI research both with children and adults. Cultural differences have been shown to affect the results of evaluation methods including survey methods such as questionnaires (Day and Evers, 1999) and the think-aloud protocol (Clemmensen et al., 2001). However there is little or no research on cultural differences when using these evaluation methods with children or on specific child focussed methods. It has been suggested that cultural differences affect how people play and interact through social networking games (Lee and Wohn, 2012) yet cultural differences in games has not been as widely researched as gender differences (De Wet and McDonald, 2006). This paper reports on a comparative analysis of children's preference between two computer games and analyses the evaluation method. It is acknowledged that many of the traditional evaluation methods that have been deployed with adults are ineffective when adopted for evaluating technology with children. Hanna *et al.* outlined adaptations that would be required for standard usability testing procedures to work with children, including how the behaviour of the evaluator may affect the children's performance (Hanna et al., 1997) but these changes do not take into account cultural differences even though research has shown that these do play a part in usability testing (Herman, 1996). Within the Child Computer Interaction domain a great deal of the early work focused upon the inclusion of children as users in usability evaluation studies of

technology (Read et al., 2001, Hoysniemi et al., 2003, Sim et al., 2006). Comparative studies emerged looking at the effectiveness of one method compared to another with children. For example the think-aloud protocol and post-task interviews were compared and the results showed no significant difference in the number of problems reported (Baauw and Markopoulos, 2004). The emphasis over the last few years has moved away from usability evaluation methods to user experience. It has been suggested that user experience is not clearly defined or well understood within the HCI community and comparing the two terms some researchers perceive usability frameworks to focus primarily on user cognition and performance while UX emphasizes non-utilitarian aspect, for example interactions (Law et al., 2008). 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* (ISO, 2010). Therefore it is important to understand these methodologies and the limitations, appropriateness in a given context, if culture can impact on the process or results in order to try and improve the methodologies. User experience is subjective and therefore cannot be captured by measuring task completion time or error rates which are traditional usability metrics. User experiences that can be captured could 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 (Carroll, 2004). 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 (Inkpen, 1997), and Malone pioneered the study of fun as an important aspect of software when developing educational learning material (Malone, 1980). Without the technology providing a positive experience children are unlikely to interact or accept it.

Several evaluation methods have emerged for use with children including Cross Age Tutoring (Edwards and Benedyk, 2007), Fun Toolkit (Read et al., 2002), Laddering (Zaman and Abeele, 2010) and This or That (Zaman, 2009). Whilst these methods have all been tested and validated with children this has not been performed across cultures. Cultural differences have been highlighted as an issue in many survey and evaluation methods across different HCI strands such as perceptions and preferences in mobile design, interface design, and subjective and objective measures of usability (Vatrapu and Suthers, 2010). Despite this these differences seem to have been neglected by the users and creators of evaluation methods with children. Evaluation methods with children have been used across the globe in many different cultures and contexts, but these potentially limit the findings to being relevant only to a specific culture they were used in. It is important therefore that research is done to identify cultural hurdles that certain evaluation methods may be susceptible to with children, to address issues that could further invalidate the results from using a specific method or in using the results to make generalised assumptions about children as a whole.

If cultural differences have been reported when using a specific method, researchers must be very careful when reporting their findings to limit the scope of the evaluation to children from a similar cultural background.

Many of these new methods rely upon the use of survey instruments or techniques however the use of survey methods with children often brings into question the validity and reliability of children responses (Horton and Read, 2008) due to their cognitive and developmental abilities and there are large differences in these abilities between children of the same age (Borgers et al., 2000). This can lead to well known issues such as satisficing, suggestibility and understanding (Read and Fine, 2005). The cognitive and developmental difference between children of different cultures are like to vary due to factors such as poverty or prosperity levels, ethnicity, differing education systems and also potentially factors such as exposure to technology. Maximizing the reliability of children's responses is vital to ensure validity and integrity of results and any recommendations or decisions made from them.

This paper aims to compare two games using the Fun Toolkit, a collection of tools created to evaluate the user experience of children before, during and after an evaluation study. Research into user experience and games has mainly focused on positive experiences such as fun or enjoyment and this will be the focus of the research presented in this paper.

2. EVALUATION METHOD

There are numerous evaluation methods that could be adopted for measuring user experiences, however, it is important that the methods had been validated with children, and therefore the Fun Toolkit was selected. This method would yield quantitative data enabling comparisons to be made.

2.1 Fun Toolkit

The Fun Toolkit comprises of a number of techniques for eliciting information from the participants. 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 1).

Figure 1: Smileyometer Rating Scale

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 to measure satisfaction (Barendregt et al., 2006) and fun (Metaxas et al., 2005) as it is easy to complete and requires no writing on behalf of the children.

The Fun Sorter requires children to rate the technology or in this instance game to a number of different constructs. The children would rank the game based upon the different constructs selecting which was best and which was the worst. An example of a completed Fun Sorter used to compare the two games in this paper is shown in Figure 2.

Fun Sorter 1

Which game was...

Question	Best	Worst
Most Fun	g	M
Easiest to Play	M	g

Figure 2: Completed Fun Sorter

The final tool is 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 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.

Would you like to play it again? 2 1 0

	Yes	Maybe	No
	✓		
	✓		

Figure 3: Again Again Table

3. METHOD

For this study, a within-subject design methodology was adopted in which the user experience of two games was evaluated using the Fun Toolkit.

3.1 Participants

In total 37 children aged between 7 and 9 participated in the study. This consisted of 20 primary school children from a single class at a UK school (referred to as SA), the children were from Year 3. SA was an inner city school comprising predominately of white, Christian children from a working class background. They took part in this study during a day of activities held at the University of Central Lancashire. They played the games individually, and there was one researcher for every two children.

For the second study, the participants were 17 primary school children from a single class at a Jordanian school in Amman (referred to as SB). The school is an independent co-educational day school offering an international education within an Arab setting. The school follows the International

Primary Curriculum (IPC) and the teaching language is English. The children were from Grade 5 (aged 8-9). In contrast to the children in SA, the children in SB were predominantly Muslim and from extremely wealthy families. The study was carried out at the School in Amman over a two day period. They played the games individually, and one researcher conducted the research over the two days.

The three researchers who took part in this study all had experience working with and conducting evaluations with children of these age groups. However one of the researchers had no experience of the Fun Toolkit, and therefore a brief training session was completed in the UK prior to the study.

3.2 Study Design

The study aimed to examine the reliability of a user experience evaluation method with children and understand if culture influences the results. The method requires the children to rate their experiences based upon a number of constructs within the Fun Toolkit.

The Smileyometer was used as a method for determining the child's overall experience of the game as it is claimed to measure feelings or experienced fun (Read, 2008).

3.3 Apparatus

The two games were all played on a HP Pavilion tx200 tablet PC, which has a touch-screen, stylus and keyboard. The children were required to play both games using the stylus to interact with the PC. The two game games were both mini games within the Purple Palace game and were judged by the researchers to be age appropriate and suitable for the evaluation. By having 2 different styles of games it was anticipated that children would have a clear preference, which could be measured for consistency.

3.3.1 Matching Game

The first game was a matching game (see Figure 4). The children simply had to select a tile and a picture would be revealed, they would select another and if they matched the score would be incremented. If they did not match the tile would resort back to its original position and the children could select another two.

Figure 4: Screenshot of matching game

3.3.2 Guessing Game

The second game was a guessing game (see Figure 5) and required the children to select one of the items, hair, eyes, nose and mouth and they would have six attempts at guessing what the person looked like behind the curtain. Once the children had selected an item from each row they would select the tick and this would tell them how many they got right. If they did not get them all right they would simply change their selections and retry.

Figure 5: Screenshot of guessing game

3.3.3 Data Capture Form

The researcher gave the children a pen and a data capture form to complete the Smileyometer, Again Again table and the Fun Sorter (see Figures 1-3).

3.5 Analysis

There were 3 forms which were partially completed and therefore were removed for analysis purposes, 2 were from SA and 1 from SB. All the remaining 34 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 fun Sorter results were coded as 1 for the highest ranked game and 0 for the lowest ranked game for each of the two constructs being examined Fun and Easiest to Play.

The final instrument in the Fun Toolkit, the Again and Again table, resulted in a score for each game with yes being coded as 2, maybe 1 and no 0. Therefore the game with the highest value would be perceived to be their preferred choice.

A cumulative score was calculated for each of the games based on an aggregation of the Smileyometer after they had played it and the responses to the Fun Sorter and Again and Again table. The maximum score a game could achieve was 5 and the lowest 1 (based on them just selecting awful from the Smileyometer).

4. RESULTS

Each of the 34 children completed the Smileyometer before and after they played each of the two games and the results are presented in Table 1.

Table 1: Mean Scores and Standard Deviation for Smileyometer

	Before		After	
	Mean	SD	Mean	SD
Guess (SA)	3.83	(0.86)	3.39	(1.09)
Match (SA)	3.83	(1.04)	4.39	(0.69)
Guess (SB)	3.88	(0.89)	3.63	(1.20)
Match (SB)	3.69	(1.14)	4.00	(1.21)

For the guessing game at SA, a Wilcoxon test revealed a significant difference between the results of the Smileyometer before and after they had played the game $Z=-2.309$, $p=0.021$, suggesting their expectations had not been met. In contrast for the Matching game their initial expectations appear to have been surpassed as the t-test revealed a significant difference between the two results $Z=-2.673$, $p=0.008$. To determine whether there was a difference in user experience between the two games a Wilcoxon test was performed, this was based upon the results of the Smileyometer after they had played the game and there was a significant difference between the two $Z=-3.140$, $p=0.002$. It is evident from the Smileyometer that children had a clear preference for the Matching game.

The results for SB were similar in the fact that there was a decrease in the children's post-hoc response to the Smileyometer for the Guessing game and an increase for the Matching game. A Wilcoxon test was performed to determine whether there was a significant difference between the two games based on the results after they played the game, $Z=-1.255$, $p=0.210$. In this case there was no

significant difference showing the children had no clear preference for one game over the other.

To establish if there was a difference between the two schools a Mann-Whitney U-test was performed on the Smileyometer after results for each game. There was no significant difference between SA and SB for the Guessing game $Z=-.766$, $p=0.444$ and similar there was no difference in the Matching game $Z=-.748$, $p=0.454$.

The next tool from the Fun Toolkit is the Again Again table and the results for this are shown in table 2 below.

Table 2: Frequency response to whether they would like to play it again

	Yes	Maybe	No
Guess (SA)	5	9	4
Match (SA)	14	4	0
Guess (SB)	3	8	5
Match (SB)	14	2	0

It is clear from the results that the majority of children in both schools had a preference for the Matching game, with no child indicating that they would not like to play it again.

The final tool is the Fun Sorter, which looked at two constructs most fun and easiest to play and the results are shown in Table 3.

Table 3: Number of children who selected a preference for a game based on the two constructs

	Guessing	Matching
Most Fun (SA)	8	10
Easiest to Play (SA)	1	17
Most Fun (SB)	4	12
Easiest to Play (SB)	9	7

In line with the other results reported for the Fun Toolkit the Fun Sorter identified a preference for the Matching game for the two constructs for SA. However, the Guessing game was judged to be easiest to play by the children in SB. This identifies a potential cultural difference between the two groups however without further exploration this cannot be confirmed as there could be other reasons for this difference.

The mean cumulative scores were calculated based on the combined results of the tools within the Fun Toolkit with a maximum value of 9 and a minimum of 1. For the Guessing game in SA, the mean cumulative score was 4.94 (SD=1.73) whilst the Matching game was 7.61 (SD=0.85). A Wilcoxon test was performed to compare the means of each game, there was a significant difference between the two $Z=-3.501$, $p<0.001$. The combined tools cumulative score within the Fun Toolkit clearly show that the Matching game is preferred over the Guessing game for the various constructs measured.

The results for SB for the Guessing game revealed a mean cumulative score of 5.31 (SD=1.89) whilst the Matching game was 7.06 (SD=1.53). A Wilcoxon test was performed to compare the means and there was a significant difference between the two $Z=-2.949$, $p=0.003$. The combined tools cumulative score within the Fun Toolkit clearly show that the Matching game is preferred over the Guessing game for the various constructs measured.

To establish if there were any cultural differences with the Fun Toolkit the cumulative scores were compared between the school in the UK and Jordan. To determine whether there was a significant difference between the scores a Mann-Whitney U-test was performed for each game. The results showed no significant difference for either the Guessing game $Z=-0.771$, $p=.441$ or the Matching game $Z=-1.017$, $p=0.309$. Based on the total scores of the Fun Toolkit this study failed to find any cultural differences.

5. CONCLUSION

Using the Fun Toolkit this study failed to find to find any cultural differences in game preference. The results from the Fun Toolkit clearly show that the children had a preference for the matching game. This preference was clear within all three tools of the Fun Toolkit. Yet there were differences reported with the construct of easiest to play, which may be associated with cultural differences. However, this brings into question whether 'easiest to play' is a good measure for preference, as in the case of SB they found the harder game most fun.

It is important that there is a coherent measure for fun in software for children, and in this study it is clear that the Fun Toolkit offers researchers and developers the opportunity to evaluate games across cultures. However it is important that further research is conducted to enhance the methods to enable decision makers to understand attributes associated with fun in software for children instead of merely a preference for one game over the other.

There are many different cultural permutations and variables that could be tested and only through this research could we truly begin to understand the inclusiveness afforded when using these methods.

6. FURTHER RESEARCH

In this study there were differences between cultures with the construct of 'Easiest to play' in the Fun Sorter. Thus further research will aim to investigate whether there are cultural differences relating to challenge and complexity of games. It is unclear whether this difference is attributed to actual gameplay or the evaluation method, as both are susceptible to culture difference.

The study yielded quantitative data from one tool other evaluations methods will be evaluated for cultural differences that will yield qualitative data. This may give more of an insight into why the child found a particular game easier to play but had a clear preference for the other.

7. REFERENCES

- Baauw, E. & Markopoulos, P. (2004). A comparison of think-aloud and post-task interview for usability testing with children. IDC 2004. Maryland: ACM Press.
- Barendregt, W., Bekker, M., Bouwhuis, D. G. & Baauw, E. (2006). Identifying usability and fun problems in a computer game during first use and some practice:. International Journal Human Computer Interaction, 64, (830-846).
- Borgers, N., Leeuw, E. D. & Hox, J. (2000). Children as Respondents in Survey Research: Cognitive Development and Response Quality. . Buletin de Methodologie Sociologique, 66, (60-75).
- Carroll, J. M. (2004). Beyond Fun. Interactions, 11, (38-40).
- Clemmensen, T., Hertzum, M., Hornbaek, K., Shi, Q. & Yammiyavar, P. (2001). Cultural cognition in usability evaluation. . Interacting with Computers, 21, (212-220).
- Day, D. & Evers, V. (1999). Questionnaire Development for Multicultural Data Collection. . International Workshop on Internationalisation of Products and Systems.
- De Wet, L. & McDonald, T. (2006). Gender and Cultural Differences 9if any!): South African school children and computer games. 5th International Conference on Computers and Games. Berlin: Springer-Verlag.
- Edwards, H. & Benedyk, R. (2007). A comparison of usability evaluation methods for child participants in a school setting. International Conference on Interaction Design and Children. ACM.
- Hanna, L., Ridsen, K. & Alexander, K. J. (1997). Guidelines for usability testing with children. Interactions, 4, (9-14).
- Herman, L. (1996). Towards Effective Usability Evaluation in Asia: Cross-Cultural Differences. . 6th Australian Conference on Computer-Human Interaction. Hamilton, New Zealand.
- Horton, M. & Read, J. C. (2008). Interactive Whiteboards in the Living Room? - Asking Children about their Technologies. 22nd British HCI Conference, 2008 Liverpool. 147-148.
- Hoysniemi, J., Hamalainen, P. & Turkki, L. (2003). Using peer tutoring in evaluating the usability of a physically interactive computer game with children. Interacting with Computers, 15, (203-225).
- Inkpen, K. (1997). Three Important Research Agendas for Educational Multimedia: Learning, Children and Gender. Educationl MultiMedia 97, 1997 Calgary. 521-526.
- Iso (2010). Ergonomics of human system interaction - Prt 210: Human-centred design for interactive system. Switzerland: International Standards Organisation.
- Law, E. L.-C., Roto, V., Vermeeren, A. P. O. S., Kort, J. & Hassenzahl, M. (2008). Towards a Shared Definition of User Experience. CHI 2008. Florence: ACM Press.
- Lee, Y.-H. & Wohn, D. Y. (2012). Are there cultural differences in how we play? Examining cultural effects on playing social network games. Computers in Human Behavior, 28, (1307-1314).
- Malone, T. W. (1980). What makes things fun to learn? Heuristics for designing instructional computer games. 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems, 1980 Palo Alto. ACM, 162-169.
- Metaxas, G., Metin, B., Schneider, J., Shapiro, G., Zhou, W. & Markopoulos, P. (2005). SCORPIODROME: an exploration in mixed reality social gaming for children. International Conference on Advances in computer entertainment technology, 2005 Valencia. ACM, 110-114.
- Read, J., Macfarlane, S. & Casey, C. (2002). Endurability, Engagement and Expectations: Measuring Children's Fun. Interaction Design and Children, 2002 Eindhoven.
- Read, J. C. (2008). Validating the Fun Toolkit: an instrument for measuring children's opinion of technology. Cognition, Technology and Work, 10, (119-128).
- Read, J. C. & Fine, K. (2005). Using Survey Methods for Design and Evaluation in Child Computer Interaction. Interact 2005. Rome.
- Read, J. C., Macfarlane, S. J. & Casey, C. (2001). Measuring the Usability of Text Input Methods for Children. HCI2001, 2001 Lille, France. Springer Verlag, 559 - 572.
- Sim, G., Macfarlane, S. & Read, J. C. (2006). All Work and No Play: Measuring Fun, Usability and Learning in Software for Children. Computers and Education, 46, (235-248).
- Vatrapu, R. & Suthers, D. (2010). Intra- and Inter-Cultural Usability in Computer-Supported Collaboration. Journal of Usability Studies, 5, (172-197).
- Zaman, B. (2009). Introduction and validation of a pairwise comparison scale for UX evaluations and benchmarking with preschoolers. Interact, 2009 Uppsala. Springer, 634-637.
- Zaman, B. & Abeeel, V. V. (2010). Laddering with Young Children in User Experience Evaluations: Theoretical Groundings and a Practical Case. IDC. Barcelona: ACM.