Evaluating Usability, Fun and Learning in Educational Software for Children

Gavin Sim
Department of Computing
University of Central Lancashire
United Kingdom
grsim@uclan.ac.uk

Stuart MacFarlane
Department of Computing
University of Central Lancashire
United Kingdom
sjmacfarlane@uclan.ac.uk

Matthew Horton
Department of Computing
University of Central Lancashire
United Kingdom
mplhorton@uclan.ac.uk

Abstract: This paper reports the findings of an investigation into the relationship between usability fun and learning in educational software designed for children. Twenty five children from an English primary school aged between 7 and 8 participated in the evaluation. A 3x3 Latin square experimental design methodology was adopted incorporating pre and post tests to measure the learning effect, observations to assess usability and fun along with a fun sorter to gauge the children's perception. The conclusions highlight the importance of fun in educational software, the difficulties in designing experiments with children and the relationship between usability and fun.

Introduction

Globally governments have set national goals and polices that identify a significant role for information and communication technology in improving educational systems and reforming curricula (Kozma & Anderson, 2002). In England targets were established for school leavers to be accredited in ICT, and all schools were to be connected to the Internet by 2002 (DFEE, 1997). Children from the age of 5 are developing basic ICT skills as a consequence of these policies, and a wide range of study material in digital format has emerged to support their learning in all subject domains. Digital content, in the form of multimedia applications, can offer greater opportunities to engage the children in learning environments, through interactive games, targeted immediate feedback and utilising sensory modalities in presenting the content. However, multimedia content may be used inappropriately, and parents and teachers are then faced with the dilemma of determining the appropriateness of the software in terms of educational content and its ability to provide a fun learning environment.

One area in which multimedia software has emerged in England is supporting children in the preparation for the SAT (Standard Attainment Task) tests at all levels of the curriculum. In England SAT tests are used at the end of Key Stages 1 (age 7), 2 (age 11) and 3 (age 14) as a means of measuring the progression and attainment of children in the national curriculum and these are seen by parents to be an important indicator of achievement. This paper examines three software packages designed to assist children in their preparation for key stage 1 SAT tests, exploring issues relating to usability, fun and learning.

Lennon and Maurer (2004) suggest that any software for children to play and learn should be an extension of the real world, media-rich, challenging, controllable, and leave room for creative activity. The software used in this study contained a diverse array of multimedia and presented the children with varied test environments but little in

the way of supplementary educational material. However, a design principle relating to assessment activities with children indicates that the questions should be embedded within the instruction itself and not presented as a separate test (Nugent, 2003). The rationale for this is that tying the assessments to the instruction reflects the premise that educational assessment does not exist in isolation, it must be aligned with curriculum and instruction if it is to support learning (Pellegrino et al. 2001). Based on these assumptions the educational benefit of the software is questionable because it fails to provide sufficient supplementary teaching material and the crucial element that may lead to learning would appear to be the feedback. Gadanidis (2004) suggests that feedback is often the weakest link in educational software offering nothing besides whether it is right or wrong. It is more effective to explain why the response is incorrect and provide them with the correct answer. This is evident from research studies into formative computer assisted assessment which has shown an increase in their understanding and learning with effective feedback (Charman & Elmes, 1998; Peat & Franklin, 2002). Therefore appropriately designed software incorporating formative feedback may have the potential to enhance the learning of children in preparation for their SAT tests. However, one unforeseen danger of adopting computer technology into education is that learning is seen as fun and entertainment (Okan, 2003).

In the traditional sense, paper based tests are deemed to be engaging but not fun (Dix, 2003). However by presenting a test in a different medium incorporating multimedia stimuli the sense of fun may be achieved. (Draper, 1999) suggests that fun is associated with playing for pleasure, and that activities should be done for there own sake through freedom of choice. By developing educational software incorporating a gaming genre this is seen as a motivational factor for children enticing them to use the software (Alessi & Trollip, 2001).

The concept of usability is an important factor in establishing whether the software will facilitate the acquisition of knowledge in educational software. ISO 9241-11(ISO, 1998) defines usability as the extent to which a product can be used by specific users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. Carroll (2004) has suggested that the concept of usability should be extended to include fun, but we prefer to stick with the ISO definition of usability, and to regard fun as a completely separate construct. (Laurillard, 2002) examines usability from a pedagogical perspective focusing on the user interface, design of the learning activities and checking whether learning objectives have been met. In educational software the interface should be intuitive and not distract the user from achieving their objective (Sim et al. 2004). The objective of educational software for children is to provide an engaging learning environment and this is usually achieved through games.

Two of the software applications in this study disguised the assessment activities within a gaming context (Fig 1. S1 & S3) whilst the other presented the material in a more formal and linear structure (S2).

Figure 1: Screenshots of the three pieces of software used in this evaluation (from left to right: S1, S2, S3)

This paper aims to evaluate the three pieces of software to investigate the relationships between usability, fun and learning in educational software for children.

It should be noted that there are two approaches to measuring usability or fun as part of a user study; one is to observe what happens, noting evidence of usability or fun as they occur during the interaction, and the other is to ask the users for their own assessments of the usability or fun in the interaction. The first of these is analogous to a doctor diagnosing patients' medical conditions by observing signs that indicate particular medical conditions, the second is analogous to the doctor diagnosing by asking the patients to describe their symptoms. Signs and symptoms are not always measures of the same things; consequently, in medicine, the best diagnosis is obtained by considering both signs and symptoms. In the same way, in a user study, measures of usability or fun taken from

observations are likely to be measuring different things to measures taken from users' own reports of usability and fun. It is unclear which of these measures best represents the aspects of fun and usability that are related to learning.

In this study we assessed 'observed usability' and 'reported usability' separately, and similarly, 'observed fun' and 'reported fun' were both assessed.

Hypotheses

A number of hypotheses were drawn up prior to the research study concerning the relationship between usability, fun and learning. It was hypothesised that fun and usability would be correlated, as any usability problems may hinder the children's ability to use the software affecting their level of fun. Further hypotheses were that learning would be correlated with both fun and usability. Within the concept of edutainment there is an implied relationship between fun and learning, and any usability problems may impede the learning process. The relationship between fun and learning may be more complicated; it is likely that increased fun could lead to more learning (this is the theoretical basis for edutainment), but it is also possible that too much fun could interfere with the learning process, and also that the amount of learning being achieved by the user might affect their enjoyment of the process.

We were interested in both 'observed' and 'reported' usability and fun, and hypothesised that 'observed usability' might be correlated with 'reported usability' and that 'observed fun' might be correlated with 'reported fun'. Also we were interested to establish which of these measures were correlated with learning.

Method

Sample

The sample consisted of 25 children of both genders, aged between 7 years 4 months and 8 years 3 months, from a Primary School (age range 4-11 years) in Lancashire, England. The whole age group from the school participated in the experiment. The sample covered the normal range of ability. Some of the children needed help with reading the questions and instructions. Not all of the children had English as their first language, but all spoke English fluently. All of the children had studied the National Curriculum for at least one year. They had completed the tests for these topics a few months earlier, so the subject matter of the software was largely familiar to them. They were about one year older than the age range for which the products were intended.

Procedure

The experimental design was piloted with a small sample from another local primary school, and as a consequence a number of questions in the pre and post tests were redesigned. The three software applications were identified as S1, S2 and S3. The design was within-subjects single factor with three conditions: S1, S2 and S3. To determine the order in which children used the three applications a 3 x 3 Latin Square was used. Product S2 presented a mixture of topics, but S1 and S3 both allowed users to choose the science topic. The experimental design ensured that each child saw different topics - 'Life Processes' on one, and 'The Solar System' on the other – in order to minimise learning effects across the software products. These particular topics were chosen because, firstly, they were treated similarly on the two products, and, secondly, they are presented in the National Curriculum as the simplest and the hardest science topics in the Key Stage 1 curriculum.

The experimental work was carried out at the school, in a room close to the children's classroom. Three similar laptops were used to conduct the experiments; they were arranged in the corners of the room to minimise distractions. One researcher sat by each laptop, and explained the tasks to the children. Each researcher was accompanied by an assistant, whose job was to note the children's reactions and engagement with the tasks. The children were withdrawn from their classroom in groups of two or three and were directed to the software. Each child came to the test as a volunteer and was given the opportunity to leave the research activity both before and during the work; none of them did so. They were all keen to take part and seemed to enjoy the experience.

Prior to using the software each child was shown its box and first screen, and was asked to indicate on a Smileyometer (Fig. 2) (Read et al. 2002) how good they thought the application was going to be. The rationale for this was that this gave a measure of expectation that could indicate whether or not the child was subsequently let down by the activity, or pleasantly surprised.

Figure 2: Smileyometer used to record children's opinions

Each child was then given a paper based pre-test based on questions found within the software to establish their prior knowledge of the subject domain. Following this the children were given instruction by one of the two researchers outlining the tasks to be performed, in each case the task was to complete a revision exercise using the software. Where children finished the task quickly, they were allowed to try some other parts of the software. The tasks were chosen to be as comparable as possible across the three products. The children were given approximately 10 minutes to use the software, after which a post-test was administered to establish any learning affect. They were then asked to rate the software for user satisfaction using a second Smileyometer to give a rating for 'actual' experience. For each of the activities the researchers and assistants recorded usability problems, facial gestures and comments to establish the level of fun. These behavioural signs are usually much more reliable than children's responses to direct questions about whether they liked something (Hanna et al. 1997). Over the course of three days every child used each of the three applications once.

A week later, the researchers returned to the school to ask the children to assess their experiences with the three pieces of software. A 'fun sorter' methodology (Read et al. 2002) was used for this final evaluation. The fun sorter required the children to rank the three products in order of preference on three separate criteria, fun, ease of use, and how good they were for learning. The method used here was to give each child a form with three spaces for each question, and some 'stickers' with pictures of the products. They were asked to rank the three products by sticking the three stickers into the spaces on the form in the appropriate order. All of the children did this successfully after a brief explanation. Most of them ranked the products differently for each criterion, indicating that they were able to distinguish between the criteria. Also they were asked to specify which of the three products they would choose, and which one they thought the class teacher would choose.

Results

Learning

Content validity was sought between the pre and post test, this is defined as the extent to which a test item actually represents the domain being measured (Salvia & Ysseldyke, 2003). For example, a questions on the pre and post test which was deemed to offer content validity is displayed in (Fig. 3), they were both assessing the same cognitive level within the subject domain. The UK National Curriculum (www.nc.uk.net) states that that children should be taught that animals, including humans, move, feed, grow, use their senses and reproduce, therefore the question was deemed appropriately aligned with the curriculum.

Figure 3: Question on the left used in the pre test and the right on the post test

However for S2 ensuring content validity and reliability for the pre and post test proved problematic as it was not possible to choose a specific area within the subject domain. The software produced 10 random questions relating to key stage 1 science curriculum incorporating aspects such as electricity, life processes, and forces and motion. Therefore during the experimental period it is probable that the children would encounter questions not related to the pre and post test. Consequently it was decided that measuring the learning effect was unfeasible for this software product in the time that we had with each child.

For S1 and S3 the children were assigned to either life process or the solar system, according to the experimental design, and the learning effect was calculated based on the difference between the pre and post test scores. A univariate analysis of variance was conducted to determine whether the learning effect was different between software S1 and S3. A significant difference was found between the two software products for life processes $F(1, 24)=6.742$ $P=0.01$. Based on the difference between the pre and post test scores the mean learning effect for S1 was 1.44 whilst S3 was 0.41 indicating that learning effect for S1 was greater than S3. This could have been attributed to the difference between the software, S3 included games that had no relationship to the curriculum whilst with S1 the activities were focussed on the curriculum.

There was no significant learning effect between S1 and S3 when analysing the results of the tests on the solar system. It is apparent that the children's knowledge regarding the solar system was limited compared to life processes, as the mean post test scores were S1 1.68 out of 6, and S3 1.88 out of 6. The difference between the pre and post test scores was to low for any conclusions to be drawn regarding any learning effect.

Observations

During the experiments each child was observed by two observers, the researcher, who concentrated on noting usability issues, and an assistant, who concentrated on observing the child, and noting indicators of enjoyment and engagement, such as comments, smiles, laughter, or positive body language, and also signs of lack of enjoyment and frustration, including, for example, sighs, and looking around the room. We considered using checklists for the observers, but decided to use free form note-taking, since pilot testing indicated that there would be a wide range of responses and issues. Scoring was done simply by counting positive issues noted for usability, and subtracting the number of negative issues; a similar algorithm was used to get a fun score for each child on each product. It is clear that these scores are very approximate; there was variability between the observers in what they wrote down, and interpreting the notes and deciding which of them were issues was a subjective process. It should be noted that the researchers and assistants were rotated throughout the experiments in order to reduce observer effects as far as possible.

Sometimes, the researchers wrote comments about fun, or the assistants wrote about usability; issues were counted up in the same way, whoever had noted them. Duplicated comments made by both observers were counted only once.

It is clear that there is a complex relationship between observed fun and usability; we had hypothesised that observed fun and observed usability would be correlated, and they were. The correlation is not strong (Spearman's rho = 0.269) but it is statistically significant (p=0.020). However, nether observed fun nor observed usability were correlated significantly with learning.

Specific usability issues with the software

By analysing the qualitative data that had been recorded, all three piece of software were found to have usability problems. Here are some specific issues that were problems for a number of children independently.

In S1 a question was displayed at the bottom of the screen, with four answers positioned to the right (Fig. 1 - left). Each answer had a coloured box around it and there was a corresponding coloured alien for each answer. These aliens popped in and out of craters in random locations and the object of the game was to feed a burger to the correct coloured alien. The aliens were the same colour as the boxes around the answers. The burger acted as a timer and appeared to be gradually eaten over a period of time, when the burger disappeared a life would be lost. Once a child had lost three lives their game was over. The game could be paused which would allow them additional time to read the question, however, none of the children appeared to notice the message informing them how to pause the game, and none of them actually paused it. Another observation recorded was that 13 of the 25 children appeared to be randomly clicking the screen in their early attempts to play this game, and 16 of the children needed an explanation on how to play the game. Following the explanation they seemed to grasp the concept of the game. The game play itself was too difficult for 10 of the children, as the aliens moved to quickly, not allowing enough time to answer the question. The game did incorporate a certain level of feedback; for example when they answered incorrectly the correct answer would be highlighted, however, only 8 of the children noticed this. In summary, the major usability issues with this game were that the procedure of answering the questions was too complex, and that answering the questions required levels of hand-eye coordination and mouse dexterity that not all of the children had. It should be remembered that the subjects here were older than the target audience for the software.

In S2 the children were given audio instruction to enter their name and then press enter (Fig. 4). They encountered difficulties from the start as 15 of them needed assistance after typing their name, as they did not know which was the enter key. Our laptops did not have a key marked 'enter'; the designers ought to have anticipated such problems and given more generally applicable instructions.

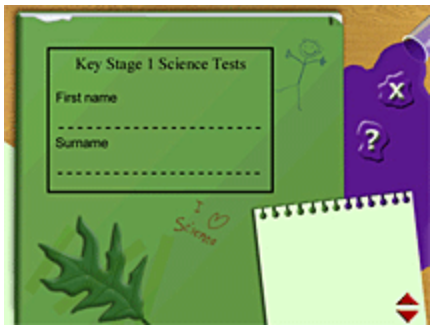

Figure 4: S2 first screen

The children were then given verbal instruction to select the revise section. Once in this section the software provided audio instructions informing them to answer the question and press the tick button which was located at the bottom right hand corner of the screen. An audio and visual representation of the question was provided by the software and all questions appeared individually on the screen. Despite the audio instructions it was noted that 10 of the children needed further assistance as they did not press the tick button after answering the question. The software allowed them to progress to the next question without comment, even though they had not recorded their answer to the previous one.

In the final software product (S3) the children had to select the appropriate topic from the list available, and then choose the revise option. They were then required to enter their name and age before beginning the questions.

Similar problems with the vocabulary used were encountered in this software as with S2, in the fact that the children did not know where the 'enter' key was located; this caused difficulty for 18 of the children. The questions appeared at the top of the screen with the four possible answers in boxes underneath (Fig. 1). A timer limited the amount of time the children could play the software, and they were allowed only to get two questions wrong before they were returned to the main menu. Three "question busters" (mechanisms to make the questions easier, such as a hint) were provided as animated icons in the top left hand corner of the screen. If the child answered a question incorrectly they would receive audio instruction to use one of the three question busters; however none of the children actually used this feature, they would just select another answer. It was apparent that the question buster feature had not been explained in a way that the users could understand. After answering a few questions the children were rewarded with a brief game. It was noted that 14 of the children found the games too difficult and failed within a few seconds, often returning a score of zero.

Evaluation by children

After all of the testing was finished, the children recorded their preferences for fun, learning, easy to use, which one the teacher would choose and the one they would choose. This was done using the 'fun sorter' method, described in the 'Method' section above.

The children's own reports of how much fun the products were to use, and of how usable they were, were correlated, with the correlation being at a very similar level to that found for observations of fun and usability. Again the correlation is not strong (Spearman's rho = 0.280) but the link is significant ($p = 0.015$).

The children were asked which software they would choose for themselves, and 13 of the children chose S3, 10 S1 and 2 S3. There was a very strong correlation between the software that the children thought was the most fun and the one that they would choose (rho = 0.658, $p<0.0005$). This shows that fun is a major criterion in the children's assessment of whether they want to use a product; this is no surprise. There was a negative correlation (rho = -0.312, $p=0.006$) between the software they perceived to be the most fun and the one that they thought their teacher would choose. There was no correlation between the children's assessment of how good a product was for learning, and whether they thought that a teacher would select it. A possible explanation of these results is that children do not see a product that is fun as being suitable for classroom use.

Conclusions

This paper has highlighted the difficulties of measuring the learning effect of educational software designed for children. The short duration of each experiment means that only a small element of the subject domain can be evaluated and it is difficult to compare different products in this way. It would be complex to replicate the experiments over a longer period of time as there are numerous variables that could not be controlled that may contribute to the children's learning, such as reading books and other supplementary teaching material. However, in two of the software, S1 and S3, there was a significant improvement in learning based on the difference between the pre and post test scores suggesting there is educational benefit in using the software.

All three software products evaluated had significant usability problems that were obvious even in a brief study such as this. Our observations showed that the children appeared to have less fun when their interactions had more usability problems. Also, their own assessments of the products for fun and usability were similarly correlated. The conclusion is that usability does matter to children, so getting it right should be a priority for designers and manufacturers.

The results also highlight the fact that the children's preference is for fun in software, which is no surprise. They clearly identified the software which presented the questions in a more formal linear manner, and which had no games, as the least fun.

A final finding was that children as young as 7-8 were able to distinguish between concepts such as usability, fun, and potential for learning. We asked them to rank the products separately on each of these criteria; they were able to

do this after a very brief explanation, and their answers showed that they were differentiating between the concepts in a consistent way.

Further Research

A series of heuristic evaluations of these pieces of software has begun, using heuristics for usability, for fun, and for educational design. These evaluations are being conducted by independent evaluators. It will be interesting to find out whether there is again a correlation between the findings for fun and for usability.

The authors are also planning a range of further investigations of evaluation methods for children's interactive products, for both usability and fun. These will include investigations of the components of the usability and fun constructs that are particularly critical for children's products, and experiments involving children as evaluators, rather than as evaluation subjects.

References

Alessi, S. M., & Trollip, S. R. (2001). *Multimedia for Learning. Methods and Development*. Massachusetts: Allyn & Bacon.
Carroll, J. M. (2004). Beyond Fun. *Interactions,* 11 (5), 38-40.
Charman, D., & Elmes, A. (1998). *Computer Based Assessment(Volume 2): Case studies in Science and Computing*. Birmingham: SEED Publications.
DFEE. (1997). *Connecting the Learning Society: National Grid for Learning. The Governments Consultation Paper*. London: Department for Educational and Employment.
Dix, A. (2003). Being Playful, Learning from Children. *Paper presented at the Interaction Design and Children, 2003,* ACM, Preston. 3-9.
Draper, S. W. (1999). Analysing fun as a candidate software requirement. *Personal Technology, 3*, 117-122.
Gadanidis, J. M. (2004). Designing learning objects for language arts pre-service teacher education in support of project-based learning environments. *Society for Information Technology & Teacher Education International Conference, 2004,* AACE, Albuquerque. 3892-3899.
Hanna, L., Risden, K., & Alexander, K. J. (1997). Guidelines for usability testing with children. *Interactions,* 4 (5), 9-14.
ISO. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability*: ISO 9241-11.
Kozma, R. B., & Anderson, R. E. (2002). Qualitative case studies of innovative pedagogical practices using ICT. *Journal of Computer Assisted Learning,* 18 (4), 387-394.
Laurillard, D. (2002). *Rethinking University Teaching: A conversational framework for the effective use of learning technologies*. London and New York: Routledge.
Lennon, J., & Maurer, H. (2004). A Child's CanDo Assistant. *World Conference on Educational Multimedia, Hypermedia & Telecommunications,2004,* AACE, Lugano.1430-1437.
Nugent, G. (2003). On-line multimedia assessment for K-4 students. *World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2003,* AACE, Hawaii. 1051-1057.
Okan, Z. (2003). Edutainment: is learning at risk. *British Journal of Educational Technology,* 34 (3), 255-264.
Peat, M., & Franklin, S. (2002). Supporting student learning: the use of computer-based formative assessment modules. *British Journal of Educational Technology,* 33 (5), 515-523.
Pellegrino, J. W., Glaser, R., & Chudowsky, N. (2001). *Knowing what students know: The science and design of educational assessment*. Washington DC: National Academy Press.
Read, J., MacFarlane, S., & Casey, C. (2002). Endurability, Engagement and Expectations: Measuring Children's Fun. *Interaction Design and Children, 2002,* ACM, Eindhoven.
Salvia, J., & Ysseldyke, J. (2003). *Assessment: In special and inclusive education.* Boston: Houghton Mifflin.
Sim, G., Horton, M., & Strong, S. (2004). Interfaces for online assessment: friend or foe? *7th HCI Educators Workshop, 2004,* Preston. 36-40.

Acknowledgement

We would like to acknowledge all the children involved in this research. A special thanks to Emanuela Mazzone, Janet Read and the students on the HCI MSc for assisting in the data collection and experimental design.