Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers & Education 46 (2006) 235–248
COMPUTERS & EDUCATION
www.elsevier.com/locate/compedu

All work and no play: Measuring fun, usability, and learning in software for children

Gavin Sim *, Stuart MacFarlane, Janet Read

Department of Computing, University of Central Lancashire, Preston PR1 2HE, UK

Abstract

This paper describes an empirical study of fun, usability, and learning in educational software. Twenty five children aged 7 and 8 from an English primary school participated. The study involved three software products that were designed to prepare children for government initiated science tests. Pre and post tests were used to measure the learning effect, and observations and survey methods were used to assess usability and fun. The findings from the study demonstrate that in this instance learning was not correlated with fun or usability, that observed fun and observed usability were correlated, and that children of this age appeared to be able to differentiate between the constructs used to describe software quality. The Fun Sorter appears to be an effective tool for evaluating products with children. The authors discuss the implications of the results, offer some thoughts on designing experiments with children, and propose some ideas for future work.

Keywords: Elementary education; Evaluation methodologies; Multimedia; Human-computer interface

1. Introduction

In the last decade, children in the developed world have become significant users of software technology. This has resulted from the investment of governments in Information and Communication

* Corresponding author. Tel.: +44 1772 895 162; fax: +44 1772 892 913.
E-mail address: grsim@uclan.ac.uk (G. Sim).

0360-1315/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2005.11.021

Technology (ICT) for schools, the fall in price of the personal computer, and the proliferation of games consoles and related dedicated platforms that have brought software and related products to the attention of children.

Software for children can be classified into three genres; enabling software, entertainment software, and educational software (Read, 2005). Enabling software includes specialist word processors, children's Internet browsers and children's art packages. Entertainment software comprises games and media products. Educational software is either linked to a published curriculum or is designed to support the mastery or assessment of a specific skill. Many authors use the term 'Edutainment' to describe educational software that has an entertainment element. The authors of this paper consider this term unhelpful, believing that a product is better defined by its main purpose. Thus, software whose main purpose is to educate would be called 'educational software' in our classification, even if it includes games and other entertainment.

It is common for educational products for young children to include games that contribute to specific skills such as Maths, however when software is intended for older users, games are more likely to be valued for their entertainment value rather than their educational merit (MacFarlane, Sparrowhawk, & Heald, 2004). Many educational software products are produced to support the 'home learning' activities that are a feature of the education system in the developed world. Parents are put under media pressure to assist their children in their education, specifically in the attainment of good grades in government initiated tests. Production of educational and leisure software is big business; currently the market for console games in the UK is worth over £1.2 billion (ITFacts, 2005). Despite this growth there is no clearly established methodology for evaluating software for children.

1.1. The national curriculum in England

The curriculum in England is presented in stages. Children are assessed at the end of Key Stage 1 (age 7), Key Stage 2 (age 11) and Key Stage 3 (age 14) by national standard attainment task (SAT) tests. These tests are used as a means of measuring the progression and attainment of children in the national curriculum (DFEE/QCA, 2000). These tests are seen by some parents to be an important indicator of achievement and ability. At the time of writing there were at least 100 commercially available software products that were built especially to support the SAT tests.

1.2. Learning, assessment and feedback

There has been much written in support of children's use of educational software at schools and at home (Chang, 2000; Kerawalla & Crook, 2002; Kong & Kwok, 2005; Smeets, 2005). Educational software that is intended to support the SAT tests is often presented as a mock test environment with practice questions being presented alongside learning activities. This method of mixing instruction and assessment supports design principles relating to assessment activities with children that suggest that questions should be embedded within the instruction itself and not presented separately (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, Glaser, & Chudowsky, 2001).

Gadanidis (2004) suggests that feedback is often the weakest link in educational software, often offering nothing more than an indication of whether an answer is right or wrong. It is more effective to explain why the response is incorrect and provide the user 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).

1.3. Fun in educational software

Malone (1980) pioneered the study of fun as an important aspect of software, and published guidelines for designing for fun (Malone, 1984). But for many years the study of fun in software was a marginal interest. In recent years there has been increasing interest in this aspect of software design (Blythe, Monk, Overbeeke, & Wright, 2003; Draper, 1999; Read, MacFarlane, & Casey, 2002). Garneau (2001) examined the activities that lead to entertainment within video games and defined 14 forms of fun. Many definitions of fun centre on emotions; and one such definition is by Carroll (2004) who suggests that things are fun when they attract, capture, and hold our attention by provoking new or unusual emotions in contexts that typically arouse none, or arouse emotions not typically aroused in given context. A problem with this definition is that it should say that the emotions should be pleasurable. Something can be engaging or captivating without necessarily being fun; Dix (2003) suggests that paper based tests are engaging but not fun. Software designers often attempt to place fun into test situations by incorporating multimedia stimuli and incorporating a gaming genre; this is seen as a motivational factor for children, enticing them to use the software (Alessi & Trollip, 2001).

The measuring of fun, especially where the users are children, is difficult (MacFarlane, Read, Höysniemi, & Markopoulos, 2003). It is possible to use a heuristic evaluation method, based, for example, on Malone's guidelines. Observational methods can also be useful, they were used in the study reported in this paper along with survey methods based on the Fun Toolkit (Read et al., 2002).

1.4. Usability in educational software

Usability is traditionally associated with work systems, and it is traditionally described using terms that relate to task driven activities within a context where the user has little discretion. 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 the present authors do not feel that this is useful. We regard usability and fun as two separate constructs, and in this paper we use the ISO definition of usability, and treat 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 the determination of whether learning objectives have been met. Usable interfaces for education need to meet standard usability requirements but should also be intuitive and not distract the user from achieving their objectives (Sim, Horton, & Strong, 2004).

It is traditional to assess usability by taking measures of the users' performance, by noting where users have difficulties with the interface, and by asking the users for their opinions of the product. It is possible to use standard usability metrics with children, but extra care needs to be taken in the interpretation of data and in the experimental set up (MacFarlane et al., 2003).

2. Hypotheses

This study was intended to investigate the relationships between usability, fun and learning in educational software for children for assessment. The main focus of the research was to examine methodological issues. There are several approaches to measure 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. Earlier studies by the authors had used both 'observed' and 'reported' fun, finding that both were informative (Read, MacFarlane, & Casey, 2001). Hypotheses for the current study were that 'observed usability' would be correlated with 'reported usability' and that 'observed fun' would be correlated with 'reported fun'. The study was also intended to determine if any of these four measures correlated with learning and whether, for children, usability and fun would be correlated.

3. Software

Appropriately designed software incorporating formative feedback may have the potential to enhance the learning of children in preparation for their SAT tests. The software evaluated in this study contained a diverse array of multimedia presented in a test environment, usually built around multiple choice questions that offered little in the way of supplementary feedback. The educational benefit of the software may be questionable because it fails to provide sufficient supplementary teaching material and feedback which is a crucial element that can lead to learning.

Three different pieces of software were used within this study (S1, S2, and S3 – see Fig. 1). One of the software applications in this study (S1) placed the assessment activities within a game context, another presented the material in a more formal and linear structure, without games (S2), and the third asked questions fairly formally, but after a few questions the users were rewarded with a separate, brief game (S3).

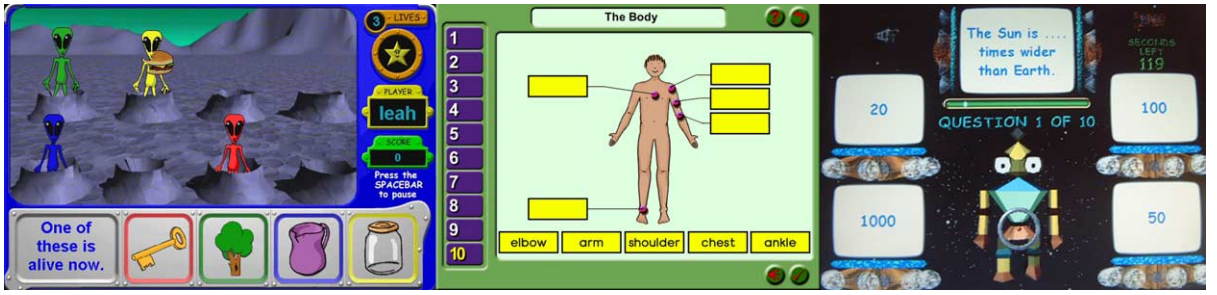

Fig. 1. Screenshots of the three pieces of software used in this evaluation (from left to right: S1, S2, S3).

In the curriculum at Key Stage 1, there are 12 different topic areas for science. These include 'Life Processes' and 'The Solar System'. Products S1 and S3 presented the material in topic areas, but Product S2 presented a randomly constructed test made up from questions from a mixture of topic areas.

Product S1 began with a menu area where the child chose the topic area. Users were then able to play various games built around answering multiple choice style questions. This research study focused on one game within the software, in which 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. The object of the game was to feed a burger to the correct coloured alien. The burger acted as a timer and appeared to be gradually eaten away over a period of time, when the burger disappeared before the question had been answered, or if the wrong answer was given, a life would be lost. Once a child had lost three lives their game was over.

Software S2 provided verbal instructions throughout with a repeat option provided. At the start, the children had to enter their name and then were presented with two options, a test or practice mode. The main difference between the two modes was that the child would receive immediate feedback after each question using the practice mode, whereas in test mode they would receive feedback at the end, and the results would be saved. For this evaluation practice mode was used. Once the child had selected practice mode they were taken to the question screen where they had to choose which question they would like to answer. The question then appeared on screen and was read out to the child. Once the child had answered the question they were required to click on the tick icon in the bottom right hand corner of the screen. The child had to go through this process until they had selected and answered all 10 questions; they would then receive a final score.

Software product S3 was similar to S1 in that the children were presented with two modes, play or revise. Within the revise option a list of subjects were available, these were topic areas within the science subject domain. The children were then asked to enter their name and age before beginning the questions. The questions appeared at the top of the screen with the four possible answers in boxes underneath. A timer limited the amount of time the children could play the software, but the time allocation increased if they scored highly on the games or answered the questions correctly. 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. After answering a few questions the children were rewarded with a brief game such as Asteroids or Arkaniod.

4. Methodology

4.1. Participants

The sample comprised 25 children (14 girls, 11 boys) aged between 7 years 4 months and 8 years 3 months. The children were all from one form class from a Primary School (age range 4–11 years)

in the North of England. The sample covered the normal range of ability with 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, and they had completed the formal tests in science a few months before the experiment. As a result of this, the subject matter of the software was largely familiar to them and at the time of using the software they were about one year older than the age range for which the products were intended.

4.2. Design of the pre and post test for learning

To determine learning, paper-based pre-tests and post tests based on questions found within the software were devised. Three different pre-tests and three different (but matched) post tests were constructed. One pair of tests was designed to test 'Life Processes' on S1 and S3, a second pair was used to test 'The Solar System' on S1 and S3 and a third pair tested the mixed selection of questions that were likely to occur in S2. It was possible to 'match' the questions to the software for S1 and S3, but for S2 the pre and post test questions did not always fit well due to the random generation of questions in this software. The questions were reviewed to guarantee they represented the subject domain ensuring content validity, this is defined as the extent to which a test item actually represents the domain being measured (Salvia & Ysseldyke, 2003).

One of the authors carried out a pilot study of the proposed experimental design with a small sample from another local primary school, and as a consequence minor changes to the layout of the questions occurred in both the pre and post tests.

4.3. Experimental design

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 3×3 Latin Square was used (Breakwell, Hammond, & Fife-Schaw, 2000). It was not possible to present a single science topic on software S2 and so the experience of the children at this software was different from that at S1 and S3. As topic areas could be controlled for S1 and S3, the experiment was designed in such a way that half the children saw 'Life Processes' on S1, and 'The Solar System' on S3 and the other half saw 'Life Processes' on S3, and 'The Solar System' on S1. These two topics were chosen because they were treated similarly within the two pieces of software, with 'Life Processes' the simplest and 'The Solar System' the hardest of the 12 in the curriculum.

4.4. Procedure

The experimental work was carried out at the school, in a room close to the children's classroom. Three similar laptops with standard mice attached were used to conduct the experiments. The children were withdrawn from their classroom in groups of two or three and were directed to one of the three laptops.

The study was carried out over three days. On the second and third day, the same children attended the experiment in a similar way to the first, but were allocated to a different piece of software. Over the course of three days every child used each of the three applications once. Each child

Fig. 2. Smileyometer used to record children's opinions.

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 the child was subsequently let down by the activity, or pleasantly surprised.

Following this, the children were given the pre-test to establish their prior knowledge of the subject area they would be presented with in the software. After completing the test the children were given instruction by the researcher outlining the tasks to be performed; in each case, the task was to complete a revision exercise using the software. The tasks were chosen to be as comparable as possible across the three products. The children were given approximately 10 min to use the software, after which the post-test was administered to establish any learning effect. 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 each child was observed by two people, the researcher concentrated on noting usability issues (mainly focussing on the screen), and one assistant concentrated on observing the child, 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. It has been suggested that these behavioural signs may be more reliable than children's responses to direct questions about whether they liked something (Hanna, Risden, & Alexander, 1997). Using checklists for the observer was considered, but it was decided to use free form note-taking, since pilot testing indicated that there would be a wide range of responses and issues. The researchers and assistants were rotated throughout the experiments in order to reduce observer effects as far as possible.

A week after the last day of testing, the researchers returned to the school to ask the children to assess their experiences with the three pieces of software. A 'fun sorter' (Read et al., 2002) was used for this final evaluation (Fig. 3). 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.

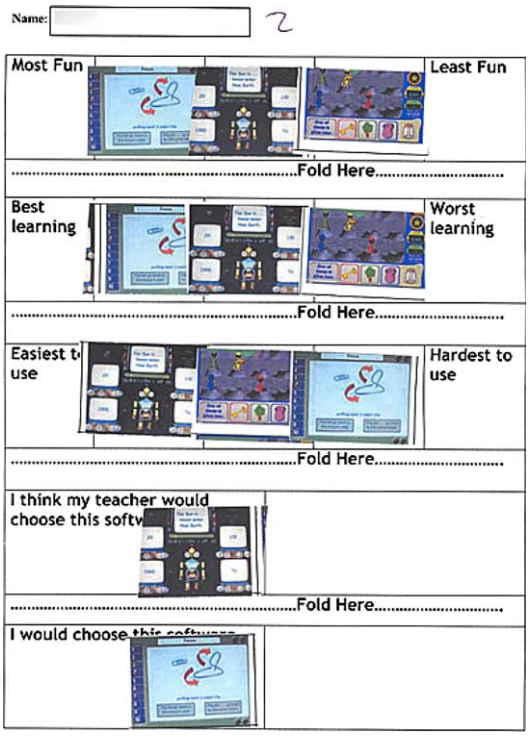

Fig. 3. Completed 'Fun Sorter'.

4.5. Analysis

4.5.1. Learning

The pre-test and the post test resulted in a numerical score out of six, where 6/6 indicated the child had got every question correct. Where a question had multiple parts partial credit was awarded as a fraction, determined by dividing the number of correct answers by the total number of parts. The learning effect was calculated based on the difference between the post and pre test scores.

4.5.2. Observed fun and usability

Scores for usability and fun were obtained from the observation data. The usability score was derived simply by counting positive issues noted for usability, and subtracting the number of negative issues. In a similar way, a fun score for each child was obtained. The researchers examined each statement documented by the observers, categorised it to either construct and unanimously had to agree whether it was a positive or negative effect. In the course of the observation, sometimes, the researchers documented instances of observed fun, and the assistants occasionally wrote about usability issues. These were included in the observed fun and observed usability scores as appropriate; care was taken that duplicated observations were counted only once.

4.5.3. Reported fun and usability

The Smileyometers were coded in an ordinal way 1–5, where 5 represented brilliant and 1 awful.

The Fun Sorters completed by the children were coded in an ordinal manner 1–3 for each of the criteria Fun, Learning and Ease of Use (Fig. 3). For example, 3 represented most fun and 1 least fun. The last two questions on the Fun Sorter Sheet, ('I think the teacher would choose this' and 'I would choose this') were scored according to how many children chose that piece of software.

5. Results and discussion

5.1. Learning

For S2, designing the pre and post tests 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 topics such as electricity, life processes, and forces and motion. Therefore during the experimental period it is probable that the children would encounter few questions related to the post test. Consequently it was decided that assessing the learning effect was unfeasible for this software product in the time that we had with each child.

For products S1 and S3, each child used one product to look at 'Life processes' and the other to look at 'The Solar System'. The former topic was straightforward, while the latter was much more difficult. Hence we decided to analyse the results separately and the learning effects are displayed in Table 1.

5.2. Observed fun and usability

The observational results for both fun and usability are presented in Table 2. An ANOVA test revealed that there was no significant difference between the three pieces of software for observed fun $F(2,72) = 2.68$, $p = 0.081$. Similarly there was no significant difference between the three pieces of software for observed usability $F(2,72) = 0.552$, $p = 0.578$.

Table 1
Mean scores for the three pieces of software for pre and the learning effect

	Pre test	Post test	Learning effect
S1 Life Process	4.33	5.78	1.45
S1 Solar System	2.12	1.68	−0.44
S3 Life Process	4.18	4.59	0.41
S3 Solar System	2.12	1.88	−0.24

Table 2
Mean scores for the three pieces of software for observations

	S1	S2	S3
Observed usability	−3.36	−4.28	−3.08
Observed fun	−0.32	0.28	1.20

The scoring method was 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. Hence only major differences could have been expected to show up as significant in this analysis.

There was a small (Spearman's $\rho = 0.269$), but statistically significant ($p = 0.020$) correlation between observed fun and observed usability. It is evident that there is a complex relationship between observed fun and usability; it was hypothesised that observed fun and observed usability would be correlated, and they were.

There was no statistically significant correlation between observations of usability or fun and the learning effect for either S1 or S3.

5.3. Reported fun and usability

5.3.1. Smileyometer results

The results from the Smileyometer data are presented in Table 3. The expectations of the children prior to using the software were all recorded between really good and brilliant for how good they thought it would be. Wilcoxon signed ranks tests were conducted to determine whether their attitude had changed after their exposure to the software. There were no significant differences between pre and post ratings for any of the products. A Friedman test comparing the ratings for the three products after use also showed no significant differences.

The majority of the children answered the questions by selecting the 'brilliant' option, so there is little variation in the answers, this enthusiasm in young children has been noted in earlier studies (Read et al., 2002).

5.3.2. Results from the 'fun sorter' ranking

Using the 'Fun Sorter' method (see Fig. 3) the children were asked to rank the three software products on a number of different criteria. Table 4 shows, for each criterion, how many children ranked each product highest.

Table 3
Mean scores for the children's responses to the Smileyometer and results of the Wilcoxon tests

Software	Before use	After use	*p*-value
S1 – Fullmarks	4.44	4.21	0.19
S2 – Europress	4.45	4.48	0.48
S3 – Active learning	4.48	4.44	0.957

Table 4
Frequency each piece of software was ranked first by the child on a number of criteria

Children's ranking	S1	S2	S3
Learning	6	7	12
Fun	13	2	10
Ease of use	11	3	11
Choose self	10	2	13
Teacher chose	3	10	12

The mean scores for fun were S1 = 2.32, S2 = 1.28 and S3 = 2.40. A Friedman test revealed a significant difference between the products, $\chi^2 = 19.52$, $p < 0.0005$. Post hoc Wilcoxon tests revealed that S2 was ranked significantly lower than both S1 and S3 ($Z = -4.054$, $p < 0.0005$), but the difference between S1 and S3 is not significant ($Z = -0.382$, $p = 0.703$). This may be attributed to the fact that S2 presented the questions in a formal linear manner and had no games.

The children were asked to rank the software in order of how good they thought they were for learning. The mean scores were 1.58 for S1, 2.02 for S2 and 2.40 for S3. There was a significant difference between the products, $\chi^2 = 8.505$, $p = 0.014$. Post hoc Wilcoxon tests revealed that the mean score for S3 was significantly higher than S1 ($Z = -2.707$, $p = 0.007$), while the mean score for S2 showed no significant difference from either S1 ($Z = -1.208$, $p = 0.227$) or S3 ($Z = -1.608$, $p = 0.108$). The children's ranking of the software which they perceived to be the best for learning is in contrast to the actual learning results where S1 scored higher than S3. One possible explanation for this is that the children do not yet have a good model of the 'good for learning' construct.

The mean scores for the children's ranking of ease of use were 2.18 for S1, 1.60 for S2 and 2.22 for S3. A Friedman test revealed a significant difference between the products, $\chi^2 = 6.143$, $p = 0.046$. In this instance the post hoc results revealed that S2 was significantly lower than S3 ($Z = -2.440$, $p = 0.015$). The difference between S1 and S2 approaches significance ($Z = -1.955$, $p = 0.051$). There is no significant difference between S1 and S3 ($Z = -0.014$, $p = 0.989$). These are similar findings to the observed usability results, which also showed S2 to have the highest number of problems, indicating that the children had probably understood the construct of 'ease of use' in a similar way to the observers.

The children's own reports of how much fun the products were to use and of how usable they found them 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 ($p = 0.015$). There was a very strong and significant 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 (not significant) between the software they perceived to be the most fun and the one that they thought their teacher would choose. There was a significant positive correlation ($\rho = 0.269$, $p = 0.020$) between the children's ranking of how good a product was for learning and whether they thought that a teacher would select it. A possible interpretation of these results is that children do not see a product that is fun as being suitable for classroom use.

There was no statistically significant correlation between reported fun or usability and the learning effect for either S1 or S3. This is a similar finding to the observational results which also revealed no correlation with learning.

6. Conclusion

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 hard to carry out 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. Within this study no correlation was found between learning and any of the measures of fun and usability.

All three software products evaluated had 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. A tentative conclusion is that usability does matter to children; so getting it right should be a priority for designers and manufacturers of software products.

The fun sorter 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. The children perceived that the teacher would choose the software based on how good it was for learning. It is evident that the fun sorter is a more effective tool than the smileyometer for evaluating software with young children.

The most interesting finding was that children as young as 7–8 appear to be able to distinguish between the concepts of ease of use, fun, and learning. They had little difficulty in completing the 'fun sorter' questionnaire. Further development of this tool should enable researchers to more easily find out children's opinions on a range of aspects of software and preferences.

7. Further research

Work has begun using a series of heuristic evaluations of these pieces of software, for usability, 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 relationship between the findings for fun and usability.

There is scope for refinement of the 'fun sorter' ranking method, but it appears to be a promising evaluation tool for use with young children, and not just for assessing fun. 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.

Following the refinement of the measuring tools it is the intention to replicate the experimental design with a different cohort of children at another primary school within the UK. The study will examine a different subject domain within the National Curriculum to ascertain whether similar results are obtained.

Acknowledgements

We thank the staff and children of English Martyrs RC Primary School, Preston. Special thanks to Emanuela Mazzone, Matthew Horton and the postgraduate students who assisted in the data collection and experimental design.

References

Alessi, S. M., & Trollip, S. R. (2001). *Multimedia for learning. Methods and development*. Massachusetts: Allyn & Bacon.

Blythe, M. A., Monk, A. F., Overbeeke, K., & Wright, P. C. (2003). *Funology: From usability to enjoyment* Human computer interaction series Dordecht: Kluwer Academic Publishers.

Breakwell, G. M., Hammond, S., & Fife-Schaw, C. (2000). *Research methods in psychology*. London: SAGE.

Carroll, J. M. (2004). Beyond fun. *Interactions, 11*(5), 38–40.

Chang, N. (2000). The teacher's attention empowering children's math learning with computers. In *Proceedings of the international conference on mathematics/science education and technology* (pp. 112–117).

Charman, D., & Elmes, A. (1998). *Computer based assessment* Case studies in science and computing(Vol. 2). Birmingham: SEED Publications.

DFEE/QCA. (2000). *The national curriculum – Handbook for primary teachers in England (key stages 1 and 2)*. London: The Stationery Office.

Dix, A. (2003). Being playful, learning from children. In *Proceedings of interaction design and children* (pp. 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. In *Proceedings of Society for Information Technology and Teacher Education international conference, Albuquerque* (pp. 3892–3899).

Garneau, P. A. (2001). Fourteen forms of fun. Gamasutra. <http://www.gamasutra.com/features/20011012/garneau_01.htm> Accessed 12.05.05.

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.

ITFacts. (2005). Consoles. <http://www.itfacts.biz/index.php?id = P3092> Accessed 12.04.05.

Kerawalla, L., & Crook, C. (2002). Children's computer use at home and at school: context and continuity. *British Educational Research Journal, 28*(6), 751–771.

Kong, S. C., & Kwok, L. F. (2005). A cognitive tool for teaching the addition/subtraction of common fractions: a model of affordances. *Computers and Education, 45*(2), 245–265.

Laurillard, D. (2002). *Rethinking university teaching: a conversational framework for the effective use of learning technologies*. London/New York: Routledge.

MacFarlane, S. J., Read, J. C., Höysniemi, J., & Markopoulos, P. (2003). Evaluating interactive products for and with children. Interact 2003, Zurich, SU: IOS Press.

MacFarlane, A., Sparrowhawk, A., & Heald, Y. (2004). Report on the educational use of games. TEEM. <http://www.teem.org.uk/publications/teem_gamesined_full.pdf> Accessed 11.04.05.

Malone, T. W. (1980). What makes things fun to learn? Heuristics for designing instructional computer games. In *Proceedings of the 3rd ACM SIGSMALL symposium and the 1st SIGPC symposium on small systems* (pp. 162–169).

Malone, T. W. (1984). Heuristics for designing enjoyable user interfaces: lessons from computer games. In J. Thomas & M. Schneider (Eds.), *Human Factors in Computer Systems*. Norwood, NJ: Ablex.

Nugent, G. (2003). On-line multimedia assessment for K-4 students. In *Proceedings of the world conference on educational multimedia, hypermedia and telecommunications* (pp. 1051–1057).

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. C. (2005). The ABC of CCI. *Interfaces, 6*(2), 8–9.

Read, J. C., MacFarlane, S. J., & Casey, C. (2001). Measuring the usability of text input methods for children. In *Proceedings of HCI2001* (pp. 559–572).

Read, J. C., MacFarlane, S. J., & Casey, C. (2002). Endurability, engagement and expectations: measuring children's fun. In *Proceedings of interaction design and children* (pp. 189–198).

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? In *Proceedings of the 7th HCI educators workshop* (pp. 36–40).

Smeets, E. (2005). Does ICT contribute to powerful learning environments in primary education. *Computers and Education, 44*(3), 343–355.