Measuring the Usability of Text Input Methods for Children

Janet Read, Stuart MacFarlane & Chris Casey

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

Tel: *+44 1772 893 276*

Fax: *+44 1772 892 913*

Email: *{jcread,sjmacfarlane,ccasey}@uclan.ac.uk*

This paper describes an experiment in which children aged between 6 and 10 entered text into a word processor using four different input methods, mouse, keyboard, speech recognition, and handwriting recognition. Several different measures of usability were made in an attempt to assess the suitability of the input methods in this situation. The paper describes and discusses the measures and their use with very young children.

Measures of effectiveness were affected by a number of different types of error that occurred during text input. Following an analysis of these errors, six types of error were identified. This analysis should help in the construction of more discriminating measures. Efficiency measures were sometimes affected by the distracting effect of novel input methods. Satisfaction measures were particularly problematic; several methods were used, with the repertory grid technique appearing the most promising.

Keywords: evaluation, children, text, speech, handwriting.

1 Introduction

This work is part of a wider project looking at alternative text input methods. Young children have relatively advanced verbal language skills yet they find spelling difficult and have difficulties constructing good written language.

Several studies have shown that the use of a word processor can assist in developing written language skills (Sturm, 1988; Newell et al., 1992). A study by Kurth (1987) suggested that although the use of a computer did not always improve the literal quality of children's writing, the high visibility of text on the screen fostered more conversation about the writing, and the spelling was improved. In his book 'Mindstorms', Papert (1980) suggests that a computer with word processing software affords the child the luxury of being able to revise and rework their ideas, and therefore becomes a powerful intellectual product.

2 Input Methods

Children in school are encouraged to use word processors to produce electronic text where possible. However, access to the software is traditionally via a QWERTY keyboard which novices find difficult to master. Hermann (1987) suggested that handling the keyboard interferes with the writing process, and if this is the case, then the use of more natural interfaces, such as speech and handwriting recognition may be desirable. There is also a case to consider whether or not children prefer to use a pointing device for text entry, this could be a touch screen or a mouse. Special keyboards are used in a few primary schools, however they have many of the same features as QWERTY keyboards and as they are not in widespread use and are considerably more expensive than the natural language interfaces they are not considered in this study. For similar reasons, touch screens are not included.

Children are able to speak fluently at an early age, and they are taught handwriting as part of the early curriculum. This suggests that the natural language technologies of speech and handwriting recognition may suit young users. Snape et al. (1997) use the term 'disobedient interface' to describe this sort of technology where a correct action by the user can result in an incorrect outcome by the software.

Studies of speech recognition have reported levels of recognition as high as 95%, but these have not been with children. O'Hare & McTear (1999) used speech recognition with secondary school age children with accuracy rates of 82%. They also reported much faster input with rates of 50wpm, compared with 10wpm at the keyboard.

Handwriting recognition is a newer technology; in order to participate, the user needs a 'pen' and a graphics tablet. An earlier experiment by the present authors (Read et al., 2000) gave an encouraging picture in relation to handwritten input with young children. In this study, children's handwriting was traced onto the machine by an adult and a Character Recognition Rate recorded. The average recognition rates exceeded 80%.

3 Design of the Experiment

The experiment was designed to compare four input methods; QWERTY keyboard, mouse clicking using an on screen alphabetic keyboard, speech input and handwritten input. These four methods were seen to be affordable to implement in a classroom and sufficiently easy to master by children of the target ages. Of interest was the comparative strengths and weaknesses of the methods, and particularly issues raised from measuring usability with children.

The experiment was carried out in a small primary school over a two-week period. Children aged between 6 and 10 were arranged in age order and systematic sampling was used to select twelve children covering all ages and abilities. Each child in the experiment did a task using each of the four input methods, but the order in which the methods were presented varied across the sample. Children did only one task a day, and at least two days separated each task. There were three parts to each task; initial orientation and/or training; copying text which was put in front of the children, and adding their own text to a story which had been introduced to them.

The text which was to be copied was constructed using words from the Key Stage 1 reading list (for Education & Employment, n.d.) (a list of words with which children aged between 6 and 8 are expected to be familiar), and was similar in structure to text which the children would have met in their school's reading scheme. Each task had its own story and this was used by all the children. The Microsoft Word grammar checking tool, was used to ensure that readability and word length were consistent across all the stories. The words used in each task were very similar. After copying each story, the children were asked to write about what happened next.

During the tasks, recordings of the children were made using a video camera.

3.1 Training

Each task required different training. The training needs had been established by discussing the interfaces with four children prior to the experiment and determining what was required. The intention was to ensure that the child was comfortable with the hardware and the software interface. A training activity was defined for each input method, and children stopped doing this activity when it was seen that they were sufficiently competent. The keyboard was familiar to all the children, so they simply typed their name and then continued with the task. The mouse interface had an alphabetic keyboard and a text display box. It did not have the extra symbols and functions which are a feature of the QWERTY keyboard as it was only being used for simple text entry. Children were shown how to select characters and were then asked to input their name. This highlighted those who were unsure about the manoeuvrability of the mouse, and they were then shown how to proceed.

The speech recognition software used was IBM (ViaVoice Millennium version 7, which had been previously trained with a young female with a local accent. The children used a headset microphone. The children tried out the software by reading some text, and were told to enunciate in a clear way, pronouncing the beginnings and endings of words clearly (Caulton, 2000). Handwriting was done using a Wacom tablet and pen, and with Paragraph Pen Office software. The software was set up for online recognition. Initially the children were encouraged to use the pen to trace over letters which were on the screen, and then they wrote a few words. They were given advice relating to the size and orientation of their script (Read et al., 2000).

3.2 Problems with the Experiment

Following the selection process, one child was identified by the class teacher as having special educational needs and it was thought that she would be unable to manage the tasks set. A second child became ill after having completed just one task;

both of these children were replaced with the adjacent child from the age ordered list.

4 What was Measured?

The measurements of the usability of the four methods were based on the standards suggested in Part11 of ISO 9241 (ISO, 2000).

Usability objectives include suitability for the task, learnability, error handling, and appropriateness (Dix et al., 1998). Some of these were inappropriate in relation to the input methods used and the user group. Learnability and appropriateness were difficult to assess given the brevity and simplicity of the tasks which were carried out. Measuring error handling would have muddled the input methods, as an attempt to correct errors created by the speech recogniser would have been difficult without using the keyboard; similarly correcting at the keyboard may have involved the mouse. The suitability of the method for the task was evaluated using Effectiveness, Efficiency and Satisfaction measures.

4.1 Effectiveness Measures

The mouse, keyboard and handwriting interfaces were all character based. The discrete characters were entered, selected, or recognised. To measure the effectiveness of these methods a Character Error Rate (CER) metric was used:

$$CER = \frac{s + d + i}{n} \times 100$$

where s = number of substitutions, d = number of deletions, i = number of insertions and n = number of characters. (A substitution counted as one error, not as a deletion plus an insertion.)

In the character systems, a decision had to be made regarding spaces. Firstly, it was noted that younger children were unlikely to put spaces in copied or composed text when using the keyboard and the mouse. Secondly, it was observed that when children put spaces in, they were as likely to put three spaces as one. Thirdly, the handwriting recogniser inserted its own single space when the pen was lifted from the tablet for a certain duration. This happened frequently with younger children, particularly when they copied text. For these reasons spaces were disregarded in the effectiveness measures.

The speech recognition was measured using 'word error rate' (WER):

$$WER = \frac{s + d + i}{n} \times 100$$

where s = number of substitutions, d = number of deletions, i = number of insertions and n = number of words. (A substitution counted as one error, not as a deletion plus an insertion.)

The CER and WER measures give a rating for the failings of the system. As effectiveness is a positive concept, a new metric, 'percentage correctness measure' (PCM) was defined:

$$PCM = 100 - (\text{either } CER \text{ or } WER)$$

4.2 Efficiency Measures

There are two measures of efficiency; these are 'characters per second' (CPS) and 'words per second' (WPS):

$$CPS = \frac{Characters\ input}{Time\ taken}$$

$$WPS = \frac{Words\ input}{Time\ taken}$$

The WPS was multiplied by the average characters per word (3.33 in this instance) to give an approximate CPS for the speech recognition.

The time taken was recorded from the start of the activity to the close. Two separate times were recorded, one for the copying exercise and one for the composing exercise. During the copying, children were generally on task for the whole time, but on four occasions, hardware and software problems caused an interruption, and this time was deducted.

During composing, children were inclined to 'pause for thought' or ask for spellings; this time was included. The rationale for this is discussed in Section 6.2.

4.3 Satisfaction

The user satisfaction was expected to be problematic with regard to the ages of the children. Satisfaction is an adult concept that correlates with a feeling that something is good enough. It is typically measured by observations and questionnaires. For adults 'Very satisfied' is used on Likert scales to refer to the best that one can get.

Watching children it soon becomes evident that 'satisfaction' is not an appropriate word for what they may be experiencing, they appear to have wider extremes of feelings than adults. Children have a different perception of the world, their cognition is less well developed which makes it difficult for them to articulate likes and dislikes (Druin et al., 1999). Evaluation techniques that work for adults may work badly or may be wholly inappropriate for children of this age.

The work of Beyer & Holtzblatt (1998) suggests that children need to have satisfaction measured in their own environment and so a suite of specially designed satisfaction metrics was developed (Read & MacFarlane, 2000).

To measure expectations, a discrete scale using a series of 'smiley' faces (Figure 1) was used (Read & MacFarlane, 2000). This was repeated after the task in order to measure the effect the activity had on the child's prior and subsequent perception of it. Selections on this scale were scored from 1 (Awful) to 5 (Brilliant).

During the task, observations of facial expressions, utterances and body language were used to establish a measure for engagement. Positive and negative signs were counted, and the balance of positive vs. negative instances was recorded (Read & MacFarlane, 2000).

When all four tasks had been completed, the children were presented with a repertory grid test (Fransella & Bannister, 1977) that had been developed with the four children who had piloted the experiment. This was used to provide comparative scores. The children were given icons representing the four methods and were asked to place them in order from best to worst in respect of the four constructs on the grid

Figure 1: The ‘smiley’ faces representing the scale.

Name of child..... Age..... Sex.....

	Best			Worst
Worked the best				
Liked the most				
Most fun				
Easiest to do				

Figure 2: The data collection grid.

(Figure 2). The method was given a score by summing the positions from 4 (best) to 1 (worst) for each attribute.

Following the repertory grid evaluation, children were asked to asked to rate each input method using a vertical ‘funometer’ (Figure 3) similar to the one developed by Risden et al. (1997). Each column on the grid measured 10cm, so a score out of ten was obtained for each method.

5 Results of the Experiments

Results for the four different text input methods are given in the tables below. The figures for Effectiveness and the first three satisfaction metrics are given as a percentage of the optimum score. The observations are absolute figures, and the Efficiency measures are in Characters per Second.

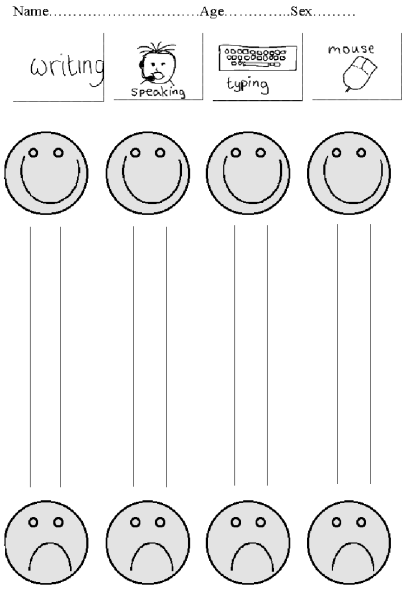

Figure 3: A prepared funometer.

	Effectiveness	Efficiency	Satisfaction			
	PCM	CPS	Rep Grid	Expec-tations	Fun-ometer	Obser-vations
Copying	90%	0.25	57%	82%	84%	20
Composing	97%	0.26				

Table 1: Keyboard input.

	Effectiveness	Efficiency	Satisfaction			
	PCM	CPS	Rep Grid	Expec-tations	Fun-ometer	Obser-vations
Copying	99%	0.14	61%	78%	77%	8
Composing	99%	0.15				

Table 2: Mouse input.

	Effectiveness	Efficiency	Satisfaction			
	PCM	0.33×WPS	Rep Grid	Expec-tations	Fun-ometer	Obser-vations
Copying	36%	5.74	67%	82%	85%	11
Composing	44%	6.26				

Table 3: Speech input.

	Effectiveness	Efficiency	Satisfaction			
	PCM	CPS	Rep Grid	Expec-tations	Fun-ometer	Obser-vations
Copying	73%	0.24	64%	88%	77%	20
Composing	86%	0.34				

Table 4: Handwriting input.

6 Observations on the Results

From this pilot study it appears that handwriting recognition closely matches keyboard entry in terms of efficiency and effectiveness. There may be considerable potential in handwriting as a text input method for young children.

It appears that the mouse is not sufficiently different from the keyboard to make it worthwhile pursuing as an alternative text input device, although there may be special cases where it is a desirable alternative.

The repertory grid and the observations both suggest that children enjoyed the handwriting and speech more than the other methods. Further work is required to establish the reasons for this.

7 Discussion of the Measures used in the Experiment

7.1 Effectiveness

Both the CER and WER are measures of errors found in the finished text. To evaluate the efficacy of these measures it is necessary to examine how each character or word is transformed into its on-screen representation. This transformation is made up of a sequence of processes which are different for ‘obedient interfaces’ and ‘disobedient interfaces’. Additionally, copying text and composing text have unique processes which contribute to the transformation.

Each process within these transformations has a capacity for error. In the table below, six different types of error are identified.

The individual processes and the errors associated with them can be seen in Figures 4 & 5. Figure 4 illustrates the errors that can occur when using an obedient interface, while Figure 5 shows the slightly different possibilities for error when using a disobedient interface.

The standard PCM measure tells us very little about which type of error has

	Example	Observations
Error 1 **Cognition error**	Child misreads a word or cannot distinguish letters.	This only happened when copying.
Error 2 **Spelling error**	Child misspells words or mispronounces a word that they know.	Rarely occurred in speech, as the children only spoke words they knew. More likely to occur in character based entry where children were unable to spell words that they wanted to use. This only happened when composing.
Error 3 **Selection error**	Child picks 'l' for 'i'.	This happened, as did 'o' for '0'. Only relevant for obedient interfaces.
Error 4 **Construction error**	Child cannot form the letter or word correctly. In handwriting, 'a' may look like 'd'. In speech, 'dragon' becomes 'dwagon'.	Only relevant with disobedient interfaces.
Error 5 **Execution error**	The child presses for too long, fails to click, or hits the adjacent character. The microphone or pen may be out of range.	This could happen with all the methods, but was less likely to occur in speech and handwriting.
Error 6 **Software induced error**	The software misrecognises the word or character.	Only a problem with disobedient interfaces.

Table 5: Classification of errors.

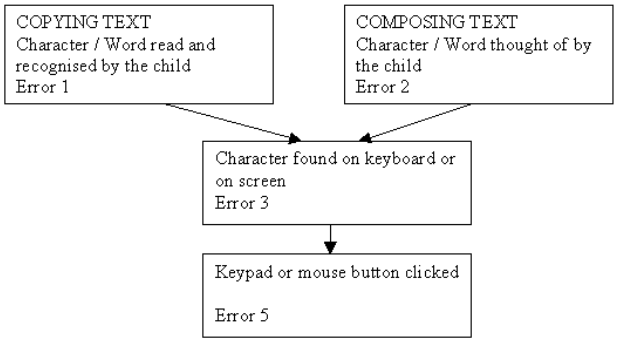

Figure 4: Processes and potential errors with obedient interfaces.

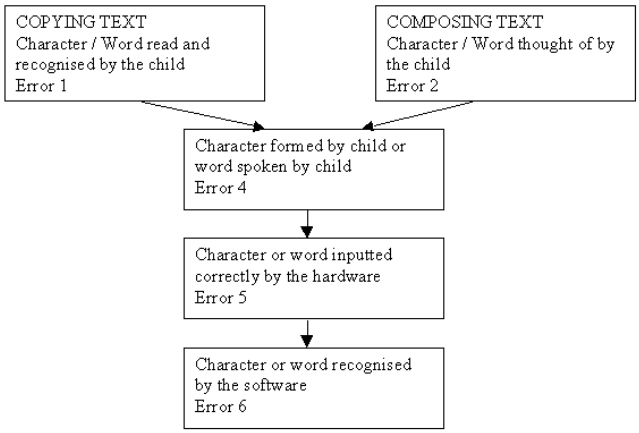

Figure 5: Processes and potential errors with disobedient interfaces.

occurred. In addition, there are errors that are averted at the last minute, which could give us information about usability. It was observed during this experiment that children sometimes intended writing something, and then changed their minds. This could be attributed to the child realising that the word they had intended was too difficult to spell, and so a substitute word was used, which masked a potential spelling error (Error 2). This didn't happen with the speech recognition software as the child was thinking and composing using words that he or she could speak (but not necessarily spell).

The speech and handwriting recognition processes are based on subcomponents of characters and words, namely strokes and phonemes. The combination of strokes or phonemes enables the recogniser to 'predict' the letter or word intended. If an error is caused by the recogniser (Error 6), then the traditional CER and WER measures will take no account of how 'close' recogniser was. The error that the speech recogniser makes in turning 'playing' into 'played in' is as bad as the one which makes 'the dragon' into 'murder again' (both are examples from this experiment). For a user, these errors are very similar, as they both require similar correction. For a researcher, trying to measure the success of an input method, the WER and CER measures cause a lot of information to be lost.

In speech recognition there have been some attempts made to address the failings of WER. A method known as phonological scoring as developed by Fisher et al. (1995) would go some way to measuring the 'closeness' of the recogniser. With this scoring method, 'playing' to 'played in' would be credited with having recognised 'play' and 'in' correctly.

Cox et al. (1998) proposed a interesting measurement based on human performance, where the question 'How much better is it than me?' would be asked,

instead of ‘How accurate is it?’ The underlying issue is to see how much better or worse the recogniser is at recognising words than a human would be. This measure was originally designed for speech recognition, but may also be useful when evaluating handwriting recognisers, as a human reader may interpret characters written by a child incorrectly, especially if they are viewed out of context.

7.2 Efficiency

The efficiency measure used was to do with time taken to input. On four occasions there were problems with the hardware or software, and in these instances, the time taken to repair the problem was subtracted from the overall time on task measurement.

Sovik & Flem (1999) describe how writing has three aspects, the motor skill, the cognitive-linguistic (thinking) process and the linguistic-semantic (spelling and grammar) process. When the child was copying text, the time taken was a measure of the motor activity associated with the task. During composition, the other aspects extended the duration of the activity. Children sometimes ‘paused for thought’ and it was noted that when children used the character based devices, they sometimes asked for spellings, or waited while they found a way of writing what they wanted to say using words which they were able to spell. This linguistic-semantic pausing was not evident during the speech input task. Cognitive-linguistic processing occurred during all the composition activities but generally it was more prolonged with the character devices. These times were included in the efficiency measures as they were seen to be part of the task of composition (Andreissen, 1992).

The children were sometimes distracted by the interface when it behaved differently than they had expected. This was particularly true of the handwriting tablet, when letters were misrecognised, the child often stopped and remarked on what had happened. It was less of a problem with the speech recogniser, partly because the children realised quite quickly that it was not showing the words they were saying, and partly because the words displayed were often incomprehensible to the children.

The distracting effect of the handwriting recogniser could possibly be masked by using ‘Lazy recognition’ wherein the text is recognised online, but is displayed after the writer has finished (Nakagawa et al., 1993). This is an interesting idea, but it would significantly change the interface, removing the potential for interactive corrections and immediate feedback.

7.3 Satisfaction

The satisfaction measures could be divided into three types; child created, child created but controlled, and observation.

There were considerable problems with some of the child created satisfaction measures. When children were presented with a Likert scale, they were generally ‘very kind’ to the activity, in that they typically thought things were going to be brilliant! The average score on the funometer was 81%. They appeared to lack the adult tendency to avoid picking the extreme values in a Likert scale. This was also the case with the discrete expectations scale, with the average being 82%.

It is likely that these measures were in part influenced by the combined effects

of several factors including the desire of the child to please, the child's perception of computers, the perceived novelty of the activity and the excitement or irritation at being taken from a class activity to take part in this experiment.

Some of these factors were ameliorated in the repertory grid scoring. However, the younger children had a limited understanding of the differences between constructs like 'Worked the best' and 'Liked the most' and so some of the ratings were likely to be unreliable. One child admitted to wanting to be fair, when it appeared that the mouse was getting a poor score!

The observations were based on negative and positive actions, these were largely beyond the child's control; and the children did not know in advance what was being observed. Assumptions had to be made about the origin of facial expressions and positive instantiations, and some children were more expressive than others.

8 Conclusions

The early intention of this study was to evaluate the usability of the four input methods. Given the small size of the user sample, it is not easy to make general statements about usability, but the results do highlight a number of areas for further study. One area is the optimisation of the recognition technologies for young children.

The inherent differences between obedient and disobedient interfaces seem to make comparisons between them difficult. There is clearly a trade off between efficiency and effectiveness. However, for children, tolerance of errors may be higher than for adults, and it may be that children prefer to use interfaces which are 'easier' and need correcting, than 'harder' interfaces which are generally more reliable.

The measures of satisfaction that were used with the children were experimental, and it seems that together the metrics can give a fairly consistent picture. It may have been illuminating to include a question / answer sheet at the end of each activity. The repertory grid test would have been improved by allowing the child to look at only one row at a time, this would have helped the child focus on one construct at a time.

Many of the considerations of error handling, efficiency and effectiveness can be applied to applications where adults rather than children are the users. Often children demonstrate exaggerated versions of adult mistakes, and they may help us to model adult behaviour.

References

Andreissen, J. E. B. (1992), Role of Representations in Reading and Writing, *in* ***EDITOR*** (ed.), *Proceedings of the Third meeting of the European Writing SIG*, Hochschulverlag, p.***PAGES***.

Beyer, H. & Holtzblatt, K. (1998), *Contextual Design: Defining Customer-centered Systems*, Morgan-Kaufmann.

Caulton, D. (2000), Can Voice Recognition Work for Computer Users? The Effects of Training and Voice Commands, *in* ***EDITORS*** (ed.), *Adjunct Proceedings of HCI'2000*, BCS, pp.77–81.

Cox, S., Linford, P., Hill, W. & Johnston, R. (1998), "Towards Speech Recognizer Assessment Using a Human Reference Standard", *Computer Speech and Language* **12**(4), 375–91.

Dix, A., Finlay, J., Abowd, G. & Beale, R. (1998), *Human–Computer Interaction*, second edition, Prentice–Hall Europe.

Druin, A., Bederson, B., Boltman, A., Miura, A., Knotts-Callaghan, D. & Platt, M. (1999), Children as Our Technology Design Partners, *in* A. Druin (ed.), *The Design of Children's Technology*, Morgan-Kaufmann, pp.51–72.

Fisher, W. M., Fiscus, J. G. & Martin, A. (1995), Further Studies in Phonological Scoring, *in* ***EDITOR*** (ed.), *Proceedings of the ARPA Spoken Language Workshop*, Morgan-Kaufmann, pp.181–6.

for Education, D. & Employment (n.d.), The National Literacy Strategy — Framework for Teaching Yr to Y6, http://www.standards.dfee.gov.uk/literacy (Accessed 7th November 2000).

Fransella, F. & Bannister, D. (1977), *A Manual for Repertory Grid Technique*, Academic Press.

Hermann, A. (1987), Research into Writing and Computers: Viewing the Gestalt, *in* ***EDITOR*** (ed.), *Proceedings of the Annual Meeting of the Modern Language Association*, ***PUBLISHER***, p.***PAGES***.

ISO (2000), "ISO 9241 International Standard. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance for Specifying and Measuring Usability". International Organization for Standardization, Genève, Switzerland. http://www.iso.ch (Accessed 6th November 2000).

Kurth, R. (1987), "Using Word Processing to Enhance Revision Strategies During Student Writing Activities", *Educational Technology* **27**(1), 13–9.

Nakagawa, M., Machii, K.and Kato, N. & Souya, T. (1993), Lazy Recognition as a Principle of Pen Interfaces, *in* S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel & T. White (eds.), *Proceedings of INTERCHI'93*, ACM Press/IOS Press, pp.89–90.

Newell, A. F., Boothe, L., Arnott, J. & Beattie, W. (1992), "Increasing Literacy Levels by the Use of Linguistic Prediction", *Child Language Teaching and Therapy* **8**(***NUMBER***), 138–87.

O'Hare, E. A. & McTear, M. F. (1999), "Speech Technology in the Secondary School Classroom, an Exploratory Study", *Computer and Education* **33**(***NUMBER***), 27–45.

Papert, S. (1980), *Mindstorms. Children, Computers, and Powerful Ideas*, Basic Books.

Read, J. C. & MacFarlane, S. J. (2000), Measuring Fun, *in* ***EDITOR*** (ed.), *Proceedings of Computers and Fun 3*, ***PUBLISHER***, p.***PAGES***.

Read, J. C., MacFarlane, S. J. & Casey, C. (2000), Postcards to Grandad — a Study of Handwriting Recognition with Young Children, *in* ***EDITORS*** (ed.), *Adjunct Proceedings of HCI'2000*, BCS, pp.51–2.

Risden, K., Hanna, E. & Kanerva, A. (1997), "Dimensions of Intrinsic Motivation in Children's Favorite Computer Activities", Poster session at the meeting of the Society for Research in Child Development, Washington, DC, USA.

Snape, L., Casey, C., MacFarlane, S. & Robertson, I. (1997), Using Speech in Multimedia Applications', *in Proceedings of TCD Conference on Multimedia*, Vol. 7(7) of *Teaching Company Directorate Seminar Proceedings*, TCD, pp.50–60. ISBN 1 901255 08 5.

Sovik, N. & Flem, A. (1999), "The Effects of Different Tasks on Children's Process and Product Variables in Writing", *Learning and Instruction* **9**(2), 167–88.

Sturm, J. (1988), "Using Computer Software Tools to Facilitate Narrative Skills", *The Clinical Connection* **11**(***NUMBER***), 6–9.

Author Index

Keyword Index