

# Validating the Fun Toolkit: an instrument for measuring children's opinions of technology

Janet C. Read

Received: 3 July 2006 / Accepted: 4 March 2007 / Published online: 22 May 2007  
© Springer-Verlag London Limited 2007

**Abstract** The paper presents the Fun Toolkit (v3), a survey instrument that has been devised to assist researchers and developers to gather opinions about technology from children. In presenting the toolkit, the paper provides a reflective look at several studies where the toolkit has been validated and considers how the Fun Toolkit should be used as well as discussing how, and in what way, the instruments contained within it should be employed. This consideration of use is one of the novel contributions of the paper. The second major contribution is the discussion based around software appeal; in which the fit between the Fun Toolkit and usability and engagement is explored. The paper concludes that the Fun Toolkit is useful, that it can be used with some confidence to gather opinions from children and that it has the potential for use for other user experiences.

**Keywords** Survey methods · Questionnaires · Fun Toolkit · Children · Evaluation

## 1 Introduction

In many studies in Child Computer Interaction (CCI), the researchers and developers seek to understand what it is that children think of the products, the applications, and the techniques, that they are developing or evaluating. There are several reasons why it is good to have this information, the first is that user satisfaction is considered to be an important element of usability (ISO/IEC 1998), suggesting

that the more satisfied a child is with a product, the more usable it will be. Secondly, establishing preferences for one product over another assists designers to make the right choices about form, content and purpose, and thirdly, there is a widely held belief that children have the right to be asked about technologies that they will use and so asking their opinions is essential.

It is possible to discover some insights into children's opinions of technology whilst interacting with them as they carry out tasks or activities. Verbalisation techniques such as think aloud, and cooperative evaluation and observational methods including the recording of facial expressions can all be employed and several studies point to the efficacy of these methods (Markopoulos and Bekker 2002; van Kesteren et al. 2003). However, verbalisation-in-use methods rely on the ability of the child to articulate opinions verbally whilst also interacting with the technology, and observation methods rely on the children being able to express their opinions in their body language. Both styles also place great demands on the researcher or developer who is required to interpret the signs and comments and formalise the responses in some way. An alternative approach is to deliberately engage with the children before or after using the technology in order to gather their first hand opinions of the product they are using. This planned engagement, where the researcher or developer is asking the child about the product in a deliberate way, can be referred to as questioning.

Questioning might take several forms; it can be informal or formal, inclusive or ad hoc, planned or unplanned. In those cases where the questioning is relatively formal, intended to be inclusive, covers a relatively large number of people, and includes some planning, it is often described as a survey (Greig and Taylor 1999). The two most common forms of surveys are questionnaires and interviews.

---

J. C. Read (✉)  
University of Central Lancashire, Preston, UK  
e-mail: jcread@uclan.ac.uk

Survey methods have long been used for gathering opinions and information from individuals, and they have a history of useful application within Human Computer Interaction (HCI) and Interaction Design. Many formal usability studies conclude with a summary of results from a questionnaire or interview, and a recent study into the use of a variety of methods with HCI practitioners in the Nordic community, highlighted survey methods as being especially prevalent and useful in usability studies (Bark et al. 2005).

Having a long history of use with adults, survey methods have also been used extensively with children with reports of their use as early as the 1890s (Bogdan and Biklen 1998). However, research about the efficacy of the different styles and usages of surveys with children is relatively scarce, and, in particular, when children are asked to contribute opinions, studies that examine the validity and reliability of the children's responses are rare (Borgers et al. 2004).

In surveys that are intended to gather opinions from children about technology, it is quite difficult to get beyond the response that all the technology is great. Often these survey studies are carried out in schools with the result that children, taken away from Maths or English, are likely to always find the novel technology that is presented to them to be a good experience!

This paper examines the validity and the usefulness of the Fun Toolkit which is a novel instrument for gathering the opinions of children about technology. The paper is loosely organized into three sections; the first section is concerned with the general use of surveys with children, the second section begins with a presentation of the Fun Toolkit as it is currently understood and the final section of the paper reflects on how the toolkit can be used and relates its use to other work in the field.

## 2 Using surveys to gather opinions

In planning and delivering any survey, there are three areas where attention is needed. The first is the consideration of the sample, the second, the mode of questioning, the third, the questions themselves (Coolican 2004). When the survey is carried out with children, the sample is often a convenience one; typically a single class or a single year group will be surveyed. In surveys with children it is important to realize that across what appears to be a well-defined sample of children there will be a wide variation of ability and skills.

Surveys can be conducted in several ways; face-to-face, by telephone, by post, by e-mail or on the Internet. Telephone, post, e-mail, and Internet surveys would be unusual with children younger than 12, but e-mail and Internet

surveys have been used with teenagers in several attitude studies (Lenhart et al. 2005; Subrahmanyam et al. 2001). With school age children, and especially in studies that are gathering opinions about technology, the survey is generally conducted face to face. Choices then have to be made about whether to use questionnaire methods or interview methods and it is necessary to decide whether each child gets to do the survey at one time, or whether each is administered the survey individually. This survey management is often dictated by circumstance, for instance if children need the questions read to them or need help with writing and if the questions relate to a recent experience with technology, then individual completion is preferred. If time precludes individual questioning, it can be possible to issue a well-designed questionnaire to a large cohort in one go.

In designing the questions for a survey Coolican (2004), proposes four general principles that specifically focus on questions. These principles are:

1. Ask for the minimum of information required for the purpose
2. Make sure questions can be answered
3. Make sure questions will be answered truthfully
4. Minimise questions that will be refused (unanswered)

The first of these principles is especially important in the design of surveys with children. It is always tempting to gather more information than is needed. In most studies all that is wanted is the first name of the child, their age and their gender. In considering the last three of Coolican's (2004) principles, careful design of the questions and piloting of the survey can ensure some reliability. In addition, a good understanding of the complexities of the question answer process can assist in the design of the survey. Breakwell (1995), describes questioning and answering as being made up of four stages:

1. Understanding and interpreting the question being asked.
2. Retrieving the relevant information from memory.
3. Integrating this information into a summarised judgement.
4. Reporting this judgement by translating it to the format of the presented response scale.

At Stage 1, children need to be able to read the question and understand what it is that the question is saying; clearly if they are poor readers or are very young, this stage is problematic. Successful completion of Stage 2 relies on the child being able to remember the details and, given that children meet many new things in a day, this needs to be assisted. Stage 3 is the point where the child decides on a response and Stage 4 is the matching of the response to the responses presented. In these latter two stages, much can

go wrong but the load on the child can be minimised by adopting specific question styles.

Several question styles of questions result in different loads at each of these stages. In a dichotomous (Yes/No) question, Stage 1 may be difficult but the following stages are made very easy. Indeed, it can be considered that answering this sort of question is too easy and it is known that children and adults have a tendency to answer ‘Yes’ in these questions (Youngman 1984). Multiple choice questions pose difficulties for children at Stages 3 and 4 as the children may not have the presented answers in their heads and may then have great difficulty choosing one of them as an answer. In addition, the ambiguity of the words in multi choice questions is known to cause difficulties (Cohen and Manion 1994). Rank Ordering is easier than multi choice but giving too many items to rank causes difficulties in Stages 3 and 4. Wilson and McLean (1994), suggest no more than five items should be ranked at a time. Rating scales, like the Likert (1932) scales and semantic differential scales (Osgood et al. 1957) also make Stages 3 and 4 easy but are more difficult to complete than dichotomous questions.

When designing questions for children, Visual Analogue Scales (VAS), which use pictorial representations that children use to assist in identifying their feelings or opinions, are often used. Although some researchers suggest that VAS can only be used with children aged around seven and over (Shields et al. 2003), studies in CCI have shown VAS to be useful for younger children, but have also noted that when these scales are used to elicit single opinions about software or hardware products, younger children are inclined to almost always indicate the highest score on the scale (Read et al. 2002a, b). This observation is expanded on further in the later stages of this paper.

Factors that particularly impact on the question—answering skills of children include developmental effects including language ability, reading age, and motor skills, as well as temperamental effects such as confidence, self-belief and the desire to please. Even in simple question styles like VAS, there is still a question to be understood, a decision to be made as to what an appropriate response will be, a decision about which visual to choose, and the physical action required to make the selection.

### 2.1 Satisficing and suggestibility

Satisficing occurs when a survey respondent (in this paper, the child) gives a more or less superficial, but nonetheless reasonable or acceptable, response to a question. Satisficing is a result of some of the steps of the question—answer process having been missed (Krosnick 1991). Satisficing is not the same as pure guessing as it generally results in a believable response and is therefore

difficult to spot. In presenting a question to a child, what the researcher or developer wants is for the respondent to go thoughtfully and carefully through all the stages of the question and answer sequence before arriving at an answer. Slippage in this process results in a degree, or level, of satisficing and this is known to be related to the motivation of the respondent, the difficulties of the task, and the cognitive abilities of the respondent (Borgers and Hox 2001). Where the respondents are children, it can be seen that, to reduce satisficing, the questions need to be especially easy to understand, and the answers need to be easy to complete.

Suggestibility concerns ‘the degree to which children’s encoding, storage, retrieval and reporting of events can be influenced by a range of social and psychological factors’ (Scullin and Ceci 2001). One of the major influences in a survey is the interviewer or researcher; even when the interviewer is trying hard not to impact on the question—answer process, when the respondents are children it is sometimes impossible to not intervene. In studies with children, the influence of the interviewer appears to be related to status; in one study, a research assistant pretended to be a police officer while asking children questions about their experience with a babysitter and the children seemed to respond to this by assuming that the nature of the experience was bad and thus the interviews yielded inaccurate and misleading results (Tobey and Goodman 1992). This was also reported in Bruck et al. (1997) who suggested that where authority figures administer surveys, the children may want to please the person administering the survey, thus providing poor results.

Satisficing and suggestibility, as well as the effects of poor reading, can be reduced by good survey design and the use of specially designed instruments that help the children respond (Shields et al. 2003).

### 2.2 Special instruments for gathering opinions from children about technology

There have been several special instruments designed for surveying with children, and some of these have been specifically created for use in studies about technology. The earliest work in this area was by (Hanna et al. 1999), who developed the first Funometer (an analogue scale to measure fun). These authors more recently reported a study into the usefulness of several commonly used rating methods (Hanna et al. 2004). This 2004 study suggested several areas for further research, in particular it reflected on the possibilities for pair wise comparisons for usability testing of products. In line with work by other researchers, the study concluded that by and large, children had high opinions of the products that they evaluated. Airey et al. (2002) have presented work with quite young children that

used tangible devices to record rankings. The children found the method easy to use, but again, as in all these studies, the authors were cautious about reading too much into the findings.

All too often, in studies in CCI, the researchers conclude that ‘children liked the product’. This is not especially helpful and the show of hands method and the Yes/No answer provides very little that is useful both for the study of the product that is being evaluated and for the further understanding of what it is that makes products appealing to children.

### 3 The Fun Toolkit (V3)

The original Fun Toolkit was developed by Read and first reported as a concept (v1) in Read and MacFarlane (2000). In 2001, the Toolkit was further developed (Read et al. 2001a, b) before being used in a research study as reported in Read et al. (2001a, b). In Read et al. (2002a, b), a theoretical underpinning for the toolkit (v2) was described and the toolkit was further reviewed in Read and MacFarlane (2006).

The toolkit in its current form comprises three instruments that can be used with children to ‘pass opinions’ on products. It is intended to be Fun, Fast, and Fair and can be used with children as young as four whilst also being acceptable for use with teenagers.

#### 3.1 The Smileyometer

The first instrument in the Fun Toolkit, and the one most used, is the Smileyometer. This is a VAS based around a 1–5 Likert scale, and uses pictorial representations as shown in Fig. 1. The Fun Toolkit is presented to the children in a horizontal row with supporting words under the faces, as recommended by Borgers et al. (2002); children are asked to tick one face.

The use of faces in these sorts of scales is not novel, there have been similar scales used in other work. In the management of postoperative pain, children have been presented with pain faces before and after surgery (Wong and Baker 1988; Bosenberg et al. 2003), and in work to find out about how children feel about their relationships with close relatives, Denham and Auerbach (1995) used a three face rating scale. The Denham and Auerbach (1995) scale uses a straight-line mouth but the Wong and Baker

(1988) scale, although having a straight smile, has eyebrows and eye features that add meaning to the faces. As with the Smileyometer, the Wong and Baker scale was also co-designed with children, but, as it was designed for pain, it would not have been easily adapted to evaluate the experience of the child with relation to computer and technology use. In the Fun Toolkit, the faces were co-designed with children aged eight and nine and this child intervention proved to be very informative as initially, the neutral state had a face with a straight-line mouth but a number of children reported that a straight-line mouth made them think of anger and so a weak smile was preferred for the neutral state.

The Smileyometer can be used before and after the child experiences the considered technology. By using it before, a measure of the expectations of the child can be gathered. Using it after a technology experience, the child is assumed to be reporting experienced feelings or experienced fun. If several technologies are being evaluated at the same time, the preferred use of the Smileyometer is to show a single one at a time for each considered product.

The key attributes of the Smileyometer are that it is easy to complete, quick to complete, requires limited reading ability, and requires no writing.

#### 3.2 The Fun Sorter

The Fun Sorter (named after the commonly found children’s toy known as the Shape Sorter) is used to compare a set of related technologies or products. Loosely based on a repertory grid (Fransella and Bannister 1977), the Fun Sorter is made up of  $n + 1$  columns (where  $n$  is the number of items being compared), and  $m + 1$  rows (where  $m$  is the number of constructs being used). One of the values of the Fun Sorter is that, by using different constructs, it can be easily used to measure more than just fun. An example of a completed Fun Sorter that was used to compare four input technologies (writing, speaking, mouse, keyboard) and presented with two constructs (worked the best, liked the most) is shown in Fig. 2.

To complete the Fun Sorter, the children need to interpret the construct and then either write a description of the technology in blank spaces, or for those children with poor reading or writing abilities, place pre-prepared picture cards on an empty grid. Having ranked the technologies or placed the cards in this way, a ranked score can be applied to each item/construct under consideration.

**Fig. 1** A Smileyometer awaiting completion



Name of child...	Age...	Sex	
	Best		Worst
Worked the best	writing	typing	speaking
Liked the most	writing	typing	speaking

**Fig. 2** A Completed Fun Sorter showing how children position the picture cards in the boxes

The use of constructs in the Fun Sorter needs special attention. Children are known to take things literally and the way they understand words cannot always be predicted; in one study it was noted that when a group of children were asked if they had been on a ‘school field trip’ they replied ‘no’ because they did not refer to the trip as a ‘school field trip.’ (Holoday and Turner-Henson 1989). In a recent study, it was noted that when children were asked how good they thought a writing activity had been, some children gave an opinion of their writing as a product, thus interpreting the question in a completely unexpected way (Read et al. 2004).

In the use of the Fun Sorter, it is recommended that, especially for younger children (<8 or 9), each construct be presented individually. Thus, in the example shown in Fig. 2, two different Fun Sorters would be prepared and the child would first fill one (with four picture cards) and then, having had that one taken away, would complete the second. Where picture cards are used, it is important to ensure that the children know what the cards represent. In the example shown in Fig. 2, the children had come across the same icons during the experimental work to which this Fun Sorter relates.

The Fun Sorter can be made so that there is no need for writing, it can be quick to complete as well as fun to complete (where stick on cards are used). Where several constructs are used, this tool becomes relatively difficult for the children to understand as each child needs to be able to read and understand the constructs and be able to see the difference. In those cases where the child cannot understand the difference and, on ranking, finds the applications or items in the same order, the child might rearrange the results to ‘suggest’ that a difference in the constructs exists even where they do not understand what the difference is.

This is a compelling reason for presenting the constructs one at a time. This tool is the most cognitively challenging of the three tools as the child may find the requirement to position and rank items according to the construct quite difficult.



### 3.3 The Again Again table

The Again Again table (after a saying made famous by one of the characters in a BBC television programme called the Teletubbies!) is a simple table that requires the child to tick either ‘yes’, ‘maybe’ or ‘no’ for each activity or product, having in each case considered the question ‘Would you like to do this again?’ The table has four columns and  $n + 1$  rows (where  $n$  is the number of activities under comparison). An example is shown in Fig. 3, where images of different products are found on the left hand side alongside three columns headed Yes, Maybe, and No. Once completed, ratings of three, two and one can be applied to the responses.

The idea for this tool comes from work in psychology that supports the idea that we are most likely to want to return to an activity that we have liked. This idea, referred to as returnance in Read et al. (2002a, b), is related to the endurance of an activity as well as the engagement felt during it.

The Again Again table cannot sensibly be used to evaluate a single product or technology. It is most useful where three or more products or activities are being compared. It needs to be presented on a single sheet after the children have experienced all the technologies and, for improved validity, the first column (showing the technologies) can be presented in different orders for different children in the sample. It is not recommended to have too many items to compare as, if there are too many rows to the Again Again table, the children get fed up filling it in (Read et al. 2002a, b).

#### Would you like to do it again?

	Yes	Maybe	No
	✓		
		✓	

**Fig. 3** An excerpt from a Completed Again Again table that was being used to compare different word processing packages

This tool is easy to complete, requires no writing, and, if pictures of applications or products are used, requires only an understanding of the single question ‘Do you want to do it again?’ and the three responses ‘Yes, No, and Maybe’. Providing there are not too many things being compared, it is especially quick to complete. The cognitive load in the Again Again table is less than in a related Fun Sorter as the child is considering each competing application or product on its own merits and is not being required to rank them one against another. This makes the tool especially well suited to the younger children.

#### 4 Evaluations of the Fun Toolkit

The Fun Toolkit tools that have been described above have been used in various combinations for a number of empirical studies by the author and by others. The successful use of the tools in these studies has demonstrated their usability and usefulness but to test the validity of the tools, several specific tests of the tools have also been carried out. Of interest in these tests of validity has been the *reliability* of answers across the tools, the use of and the *understanding* of constructs in the Fun Sorter, and the *effectiveness* of the tools for the different *ages* of children.

##### 4.1 Reliability across the tools

This has been tested in a number of studies that have compared one tool against another. In a study reported in Read and MacFarlane (2006), the Again Again table and the Smileometer, when used after the child had experienced the technology, were shown to be strongly correlated. This study used a cross ability sample of 24 children aged eight and nine, and the context of use was an evaluation of computer games. Sixty pairs of results were included in this study and the correlation between these pairs was very high (Spearman’s  $\rho = 0.780$ ,  $p < 0.0005$ ). The same paper reported a correlation between the Fun Sorter and the Again Again table when the construct of interest of the Fun Sorter was Fun. In this work, 15 children aged seven and eight used three different writing interfaces and again, there was a high correlation (Spearman’s  $\rho = 0.526$ ,  $p < 0.0005$ ). The study also reported a much weaker, non-significant correlation between the Fun Sorter when measuring ease of use and the Again Again table suggesting that these two things are not measuring the same thing.

##### 4.2 Understanding what is being measured

From the work reported in Read et al. (2002a, b), given that the correlations across the Fun Sorter and the Again Again table differed according to the construct being applied, it is

reasonable to suppose that the children are aware of the differences between constructs. However, this understanding is heavily influenced by the ability of the child and their age. In order to further understand how children understand the constructs in the Fun Sorter, data from a study reported in Read et al. (2001a, b) is examined here. This study involved only a small number (13) of children aged between six and nine but as each had been presented with four different computer input methods to evaluate, there was a large amount of attitudinal data. Once the children had met the four input technologies, they were singly presented with a Fun Sorter that showed four constructs (always in the same order), these were ‘*Worked the best*’ (W), ‘*Liked the most*’ (L), ‘*Most Fun*’, (F) and ‘*Easiest to use*’ (E). Completion of the grid was made easy for the children by presenting them with pictures that represented the four input methods on small pieces of card and asking them to lay the cards on an empty grid (using the same method as shown in Fig. 2).

Each child produced four sets of data from the Fun Sorter; these figures are shown in Table 1. The data for Child 3 (C3) for the Pen indicates that this child had placed the pen in the third place out of four for Worked the best, Liked the most, and Easiest to use, and had placed it second place out of four for Most Fun.

In this grid, the values in bold represents those results where the children showed no variability in the scores across the four constructs. The values in italic represents a small (1 point) shift in variability across the four constructs. Inspection of this table shows some that 29% of the ratings had no variability across the constructs, 31% had a one point variability, and only 40% varied by more than one. Given that some of the constructs were quite similar, (‘Liked the Most’ and ‘Most Fun’) this is not especially surprising but ‘Worked the best’ and ‘Easiest to use’ were expected to provide some opposite scores which were not all than evident; for instance, the speech was very easy to use but worked badly. Only children C2, C5 and C11 made this connection. Table 2 shows the overall correlations across the constructs.

A Spearman  $\rho$  showed that there was no difference between the constructs. These correlations, and their  $p$  values (not reliable), are shown in Table 2. It may well be the case that this lack of variability was caused by the effects of children C1, C8, and C12, all of whom were 6 year olds, and all of whom showed little variability.

##### 4.3 Age effects

Considering that there appeared to be a difference between younger and older children with respect to construct behaviour, differences in the recording of fun were also expected. To test this theory, a trial was designed to look at

**Table 1** Ratings from the two tools

	A G E	Pen	Keyboard	Speech	Mouse
		W L F E	W L F E	W L F E	W L F E
C1	6	3 3 3 3	4 4 4 4	2 1 2 2	1 2 1 1
C2	6	1 1 1 3	2 2 2 2	4 4 3 1	3 3 4 4
C3	8	2 2 3 2	3 3 2 3	1 1 1 1	4 4 4 4
C4	9	4 4 4 4	2 3 1 3	1 2 3 1	3 1 2 2
C5	7	1 2 2 2	2 4 4 3	4 1 3 1	3 3 1 4
C6	7	1 2 1 3	4 3 3 2	2 1 2 1	3 4 4 4
C7	7	1 4 2 3	3 1 1 1	4 3 4 4	2 2 3 2
C8	6	1 1 1 1	2 2 2 2	4 3 3 3	3 4 4 4
C9	8	4 3 1 3	2 1 4 1	1 2 3 2	3 4 2 4
C10	7	2 2 2 2	1 1 1 1	3 3 3 3	4 4 4 4
C11	8	3 1 1 2	4 2 2 1	1 3 3 3	2 4 4 4
C12	6	1 1 1 1	3 3 2 3	2 2 3 2	4 4 4 4
C13	7	2 4 3 4	3 2 2 2	1 1 1 1	4 3 4 3

**Table 2** Correlations between constructs

	Worked	Liked	Fun
Liked	Rho = 0.477 <i>p</i> = 0.000		
Fun	Rho = 0.477 <i>p</i> = 0.000	Rho = 0.615 <i>p</i> = 0.000	
Easiest	Rho = 0.415 <i>p</i> = 0.000	Rho = 0.800 <i>p</i> = 0.000	Rho = 0.508 <i>p</i> = 0.000

age effects during a web-site design project with children. In this study, 53 children aged between 8 and 10 used Smileyometers before and after a website design event. The Smileyometer was scored from 1 to 5 (where 1 represented the lowest rating and 5 represented the highest rating) and it was interesting to note that, as shown in Table 3, the average score for the predicted expectations of the 9/10 year olds was significantly lower than the average score for the 8/9 year olds although each reported similar ratings for the actual event once it had taken place.

Similar results were reported in Read and MacFarlane (2006) where 47 children aged between seven and nine and 26 children aged 12 and 13 were asked to complete Smi-

**Table 3** Average scores for expectations before and after

	Before	After
9–10	3.9	4.6
8–9	4.4	4.6

leymeters after completing game related tasks and in this instance there was a significant ( $U = 5786.5, p = 0.006$ ) difference between the two age groups (the mean for the older children was 3.5 and for the younger children 3.9). The older children showed a much higher degree of variability across the Smileyometers.

### 5 Related work

As indicated earlier in the paper, the Fun Toolkit has been used in many studies by the author but also in several studies by other authors. In Barendregt et al. (2006), Smileyometers were used by 25 children aged between five and seven after a first experience with a game and after their last experience. The authors found that the paired use of Smileyometers in this way showed that children appreciated the game more after the last session than they had after the first session ( $Z = -2.46$  based on negative ranks). The use of the paired Smileyometers to track changing satisfaction levels over time, as opposed to comparing before and after scores, seems worthwhile.

Metaxas et al. (2005) used paired Smileyometers to measure expected and experienced fun by asking 12 children aged between 8 and 12 to rate a mixed reality game before and after play. Before the game play, the children were given a description of the game and when it was clear that they understood the game they were asked to complete the first Smileyometer. Using an idea that is proposed in the original toolkit, the authors also asked the children to note what they had liked about the game straight after it was played, and then, some time later, they were asked the same question to see what had ‘stuck’ in their minds. This is referred to as returnance in the original work by Read et al. (2002a, b) and is considered to be one measure of the endurance of the game. In the current version of the Fun Toolkit, endurance is measured primarily by the use of the Again Again table. In the study by Metaxas et al. (2005), the paired use of Smileyometers indicated that the children’s expectations were high but also that they were met by the game. A month after the game, the children remembered almost as much as they had before indicating high levels of remembrance. The Metaxas et al. (2005) study also used a variation on the Again Again tool by asking the children if they wanted to play the game again. Unsurprisingly, all the children said yes and this perhaps indicates that yes, the game was engaging, but also that the use of Again Again without comparative items is limited—used in this way it is little more than a show of hands.

MacFarlane et al. (2005) used Fun Sorters and Smileyometers in a study that compared Fun and Ease of Use in educational software. This study demonstrated differences

in the understanding of Fun and Usability as reported by the children.

## 6 Discussion

What is the Fun Toolkit measuring? As indicated earlier in this paper, the Fun Toolkit was originally intended to be measuring some variation of user satisfaction which was loosely referred to as fun. However, changing constructs, evidence from poor and strong correlations, and the use of the toolkit in different ways indicates that there is rather more to the toolkit than simply fun. In the study reported in Read and MacFarlane (2006) it was claimed that it is the fun of a product that determines whether children will want to use it again and certainly, if this is the case, the relationship between fun, usability, ease of use and user selection needs to be further examined.

There have been several attempts to investigate how fun, usability and user satisfaction are related. Carroll and Thomas (1988), urged the HCI population to ‘take fun seriously’. One of the major contributions of this work was to highlight the difficulties inherent in the vocabulary that is used to describe software. This vocabulary includes phrases like, ‘easy to learn’, ‘easy to use’, ‘fun’, ‘productive’ and ‘friendly’; these phrases are quite similar to the phrases used in the Fun Sorter. The particular focus of the Carroll and Thomas (1988) paper is the conflict between fun and ease of use and the authors conclude that one major problem with fun is in measuring it; a piece of software can be designed to be fun, but there is no guarantee that the user will experience fun. In measuring fun, researchers in CCI have an advantage in that, by and large, the products they are presenting to children are fun, and the children, despite any limitations of the products under investigation, will report having fun. This brings a complication into the measuring of fun that is inherent in all work with children; as reported in Read et al. (2002a, b), in Sim et al. (2006) and Read and MacFarlane (2006), children do find almost all things fun.

From being considered to be something other than usability, fun is now seen as a possible part of usability. In 2004, Carroll revisited fun and usability, urging researchers to ‘construct a broader, more encompassing concept of ‘usability’, one that incorporates ‘fun’ and other significant aspects of human interaction with technology, rather than settling for the primitive caricature of usability as synonymous with simplicity and ease, and regarding fun (and other aspects of the user experience) as something beyond or aside of usability.’ (Carroll 2004, p. 39).

It could, therefore, be argued that usability is the wrong place to start when determining how good something is. By its nature, usability focuses on the negative aspects of

technology, and on those things that cause poor usability, but in traditional usability measures, these do not seem to be able to be counterbalanced with things that cause good usability. Thus it may be that ease of use is about the absence of bad things and fun is about the presence of good things.

A useful model that is along these lines is the one proposed by Hassenzahl et al. (2000) which defines a new metric ‘software appeal’ which is in part dependent on ergonomic aspects like simplicity and controllability and is also affected by more hedonic aspects, these include novelty and originality. The Hassenzahl et al. (2000) model, shown here in Fig. 4, demonstrates how these relate.

In their study, which was based on experiences with adult users, Hassenzahl et al. (2000) conclude that software appeal is equally contributed to by hedonic and ergonomic quality, implying, that for adult users, these things have equal significance. Interestingly, the study also showed that there was no difference in overall software appeal before and after use, but that the individual components changed with the ‘before’ score being more closely related to the hedonic quality and therefore the ‘after’ score being more related to ergonomic aspects.

When this model is considered in the light of the Fun Toolkit several interesting parallels and differences emerge. When used with several different constructs, e.g. ‘Ease of Use’ and ‘Fun’, the Fun Sorter can be used to measure both the Ergonomic and Hedonic qualities of the software or of the application. The Smileyometer, when used before an application is measuring expectation which for children, as for adults, will also be more of a hedonic quality. After use, if the child has experienced poor usability, the score may plummet, indicating a shift in emphasis to the more ergonomic qualities. This could be tested by introducing children to technologies that were deliberately poor in respect of usability. The Again Again table is measuring software appeal as determined by the child. That this is more likely to be about hedonic than

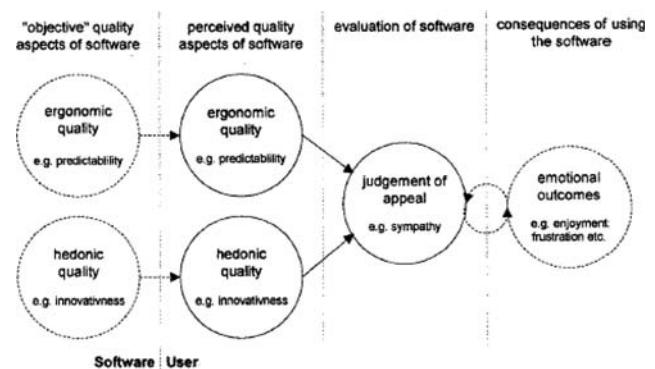


Fig. 4 Software appeal



ergonomic experiences is evidenced in the study from Read and MacFarlane (2006) that links Again Again scores to the experience of fun rather than to the experience of usability.

## 7 Conclusion

Because any survey is, by definition, designed, it will always be restrictive. Researchers and developers of interactive products are generally not specialists in survey design and so invariably produce questions and suggested answers that are far from perfect. It is common, and not unexpected, to find that in many studies, the questions are asked in such a way that the answers are invariably the ones the survey designers wanted to hear. Decisions to ask children questions in positive ways (which is promoted as good practice by psychologists and educators), for example ‘How much fun was that?’ and ‘How easy was that to use?’ may result in much different responses to questions like ‘How bad was that?’ or ‘How boring was that?’ Further work is therefore needed to investigate the effect of these negative questions, such as those commonly included in adult surveys to test the validity of the questioning instrument.

Other interesting areas for further research include the effect of gender on survey results and the effect of the type of technology under review. It is the author’s belief that the more novel the technology is, the more unpredictable the children’s responses will be, but this has not yet been tested. In addition, there are several studies to be done about the stability of the children’s opinions over time and the effects of prolonged exposure to the technologies or products under review.

Cynics may dismiss the opinions of children, considering them to be almost not worth gathering; the alternatives, observing children’s behaviours, using body indicators, like sweat and heart rate, seem no more attractive. If researchers want to know what children think of products, easy, fun to use instruments have a value.

The Fun Toolkit, as presented here is Fun, Fast and Fair. The tools in the Fun Toolkit gather only what is needed, are easy to answer, encourage truthful completion, use few written words and are easily adapted. Used carefully they can provide useful information for researchers and developers about children’s preferences for different technologies.

**Acknowledgments** Hundreds of children have freely shared their opinions and views, they are acknowledged here for their willingness to provide windows into their worlds, for moments of amusement, and for the many special insights they have provided. Also, thanks to Dr Stuart MacFarlane who provided some assistance with the statistics in this paper.

## References

- Airey S, Plowman L, Connolly D, Luckin R (2002) Rating children’s enjoyment of toys, games and media. In: 3rd world congress of international toy research on toys, Games and Media, London
- Barendregt W, Bekker MM, Bouwhuis DG, Baauw E (2006) Identifying usability and fun problems in a computer game during first use and after some practice. *Int J Human Comput Stud* 64:830–846
- Bark I, Folstad A, Gulliksen J (2005) Use and usefulness of HCI methods: results from an exploratory study among nordic HCI practitioners, HCI 2005. Springer, Edinburgh
- Bogdan RC, Biklen SK (1998) *Qualitative research for education: an introduction to theory and methods*. Allyn and Bacon, Boston
- Borgers N, Hox J (2001) Item non response in questionnaire research with children. *J Official Stat* 17(2):321–335
- Borgers N, Hox J, Sikkel D (2002) Response quality in research with children and adolescents: the effect of labelled response opinions and vague quantifiers. *Int J Public Opin Res* 15(1):83–94
- Borgers N, Hox J, Sikkel D (2004) Response effects in surveys on children and adolescents: the effect of number of response options, negative wording, and neutral mid-point. *Qual Quant* 38(1):17–33
- Bosenberg A, Thomas J, Lopez T, Kokinsky E, Larsson IE (2003) Validation of a six-graded faces scale for evaluation of postoperative pain in children. *Paediatr Anaesth* 13:708–713
- Breakwell G (1995) *Research methods in psychology*. SAGE Publications, London
- Bruck M, Ceci SJ, Melnyk L (1997) External and internal sources of variation in the creation of false reports in children. *Learn Individ Differ* 9(4):269–316
- Carroll JM (2004) Beyond fun. *Interactions* 2:38–40
- Carroll JM, Thomas JC (1988) Fun. *SIGCHI Bull* 19(3):21–24
- Cohen L, Manion L (1994) *Research methods in education*. Routledge, London
- Coolican H (2004) *Research methods and statistics in psychology*. Hodder and Stoughton, Abingdon
- Denham SA, Auerbach S (1995) ‘Mother–child dialogue about emotions and pre-schoolers’ emotional competence. *Genet Soc Gen Psychol Monogr* 12(3):311–327
- Fransella F, Bannister D (1977) *A manual for repertory grid technique*. Academic, London
- Greig A, Taylor A (1999) *Doing research with children*. Sage, London
- Hanna E, Ridsen K, Czerwinski M, Alexander KJ (1999) The role of usability research in designing children’s computer products. Druin A (ed) *The design of children’s technology*. Morgan Kaufmann, San Francisco, pp 4–26
- Hanna L, Neapolitan D, Ridsen K (2004) Evaluating computer game concepts with children. IDC2004, ACM Press, Maryland
- Hassenzahl M, Platz A, Burmester M, Lehner K (2000) Hedonic and ergonomic quality aspects determine a software’s appeal. CHI2000, ACM Press The Hague, Amsterdam
- Holoday B, Turner-Henson A (1989) Response effects in surveys with school-age children. *Nurs Res (methodological corner)* 38:248–250
- ISO/IEC (1998) 9241—14 Ergonomic requirements for office work with visual display terminals (VDTs)
- van Kesteren I, Bekker MM, Vermeeren APOS, Lloyd P (2003) Assessing usability evaluation methods on their effectiveness to elicit verbal information from children subjects. IDC2003, ACM Press, Preston
- Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl Cogn Psychol* 5:213–236

- Lenhart A, Madden M, Hitlin P (2005) Teens and technology: youth are leading the transition to a fully wired and mobile nation. PEW Internet and American Life, Washington DC, p 57
- Likert R (1932) A technique for the measurement of attitudes. *Archives of Psychology* 140
- MacFarlane SJ, Sim G, Horton M (2005) Assessing usability and fun in educational software. IDC2005, Co, ACM Press, Boulder
- Markopoulos P, Bekker M (2002) Usability testing with children subjects: comparing usability testing methods. *Interaction design and children*, Eindhoven
- Metaxas G, Metin B, Schneider J, Shapiro G, Zhou W, Markopoulos P (2005) SCORPIODROME: an exploration in mixed reality social gaming for children. ACE 2005, ACM Press, Valencia
- Osgood CE, Suci GJ, Tannenbaum PH (1957) The measurement of meaning. University of Illinois Press, Urbana
- Read JC, MacFarlane SJ (2000) Measuring fun—usability testing for children. *Computers and Fun 3*, BCS HCI Group, York, England
- Read JC, MacFarlane SJ (2006) Using the Fun Toolkit and other survey methods to gather opinions in child computer interaction. *Interaction Design and Children*, IDC2006, ACM Press, Tampere
- Read JC, MacFarlane SJ, Casey C (2001a) Expectations and durability—measuring fun. *Computers and fun 4*, York, England
- Read JC, MacFarlane SJ, Casey C (2001b) Measuring the usability of text input methods for children, HCI2001. Springer, Lille
- Read JC, Gregory P, MacFarlane SJ, McManus B, Gray P, Patel R (2002a) An investigation of participatory design with children—informant, balanced and facilitated design. *Interaction Design and Children*, Shaker Publishing, Eindhoven
- Read JC, MacFarlane SJ, Casey C (2002b) Endurability, engagement and expectations: measuring children's fun. *Interaction Design and Children*, Shaker Publishing, Eindhoven
- Read JC, MacFarlane SJ, Horton M (2004) The usability of handwriting recognition for writing in the primary classroom, HCI 2004. Springer, Leeds
- Scullin MH, Ceci SJ (2001) A suggestibility scale for children. *Person Individ Differ* 30:843–856
- Shields BJ, Palermo TM, Powers JD, Grewe SD, Smith GA (2003) Predictors of a child's ability to use a visual analogue scale. *Child Care Health Dev* 29(4):281–290
- Sim G, MacFarlane SJ, Read JC (2006) All work and no play: measuring fun, usability, and learning in software for children. *Comput Edu* 46(3):235–248
- Subrahmanyam K, Greenfield P, Kraut R, Gross E (2001) The impact of computer use on children's and adolescents' development. *Appl Dev Psychol* 22:7–30
- Tobey A, Goodman G (1992) Children's eyewitness memory: effects of participation and forensic context. *Child Abuse Neglect* 16:807–821
- Wilson N, McLean S (1994) Questionnaire design: a practical introduction. Co Antrim, University of Ulster Press, Newton Abbey, pp 94–120
- Wong DL, Baker CM (1988) Pain in children: comparison of assessment scales. *Pediatr Nurse* 14:9–17
- Youngman MB (1984) Designing questionnaires. In: Bell J, Bush T, Fox A, Goodey J, Goulding S (eds) *Conducting small scale investigations in education management*. Harper and Row, London, pp 156–176