TABLE OF CONTENTS

OBJECTIVES OF THE COURSE

The goals of participants in this course will be:

- To discover how to interpret outputs from an evaluation session with children
- To become familiar with some evaluation methods that have proved useful with child users.
- To discover some of the pitfalls in planning and carrying out evaluations with children and discover tricks and tips to eliminate most of these
- To partially appreciate the challenges and rewards of working with children as evaluators.

INTRODUCTION

Many evaluation methods have been devised, documented, and used successfully in situations where the intended users are adults. These methods include observational and survey methods that involve representative users, and methods such as walkthrough that do not involve users directly but assume a knowledge and understanding of their actions. Most of these methods need special approaches, if they can be used at all, when the intended users are children. In addition, most evaluation methods used with adults are concerned primarily with the usability of a product; with children, usability remains important, but evaluating fun is likely to also be a major concern.

We have used a variety of evaluation methods and situations involving children, evaluating both usability and fun. Our aim in this new tutorial is to pass on some of our experience of what works and what doesn't to practitioners and researchers who may need to carry out evaluations of interactive products for children.

The course is aimed at developers of interactive products for children who may need to carry out evaluation studies. Both experienced evaluators and people new to the field will find the tutorial useful.

CHILDREN

The key to success of any product design is an understanding of a target group. This chapter briefly describes the major child development theories and then goes on to discuss those characteristics of children that particularly impact on the choice and usefulness of evaluation methods. We conclude with a discussion of some of the roles that children assume when they interact with technology.

Basic Child Development

Children are an interesting user group. There are many texts on child development, and it would be impossible to adequately cover all the theories in this set of tutorial notes. Instead, we summarize the main findings and emphasise those points that impact on the evaluation of interactive products. Child development tends to focus on the five 'features' of children that 'change' with age. These are:

Physical development

Social development

Emotional development

Intellectual development

Language development

It is common to bracket children into discreet age groups. These are sometimes called infancy (0 –2), early childhood (3 – 8), later childhood (9 – 12) and adolescence (13 – 18). Some experts sub divide early childhood into a group aged 3 – 4 and a group aged 5 – 8. Others take the 11 and 12 year olds out of the later childhood stage.

Piaget's (1929) stages of cognitive development can be useful to identify the key stages of intellectual and language development.

Stage	Ages	Key points for evaluation of products
Sensorimotor	Birth to 2 years	
Preconceptual thought	2 - 4	
Intuitive thought	4 – 7 years	Children can use symbols and words. Children can distinguish reality from fantasy. In the latter part they can take into account the viewpoint of others.
Concrete operations	7 – 11 years	Children can classify things, understand the notion of reversibility and conservation. Can think logically but not abstract
Formal operations	11 years on	Thinking is about ideas, they can consider various solutions without having to act them all out – can deal with hypothetical situations

Table 1

Work by Erikson (1950) on social and emotional development suggests that around aged 4 – 5, children become competitive and start to prefer sex-appropriate activities. At this age they begin to feel both responsibility and guilt. They are not able to use their own initiative until aged around 7, when they also learn to follow rules. By age 13 (adolescence), children are aware of who they are but may experience minor delinquency, self-doubt and rebellion. This makes this group a difficult one for user studies!

The Temperament of Children

Individual children have different temperaments. These can have a significant effect on the usefulness and validity of a user study. A child's temperament is unlikely to change as he develops; Chess and Thomas (1996) have identified nine different temperamental dimensions.

Activity Level. Different children work at different paces.

Tip: Allow children time to complete tasks – build in slack.

Distractability. This relates to the degree of concentration that the child has. Hanna, Risden and Alexander (1997) consider that in a usability study, pre-school children can concentrate for about 30 minutes.

Tip: Keep sessions as short as possible – look for signs of distaction

Sensory threshold. How sensitive the child is to noise etc.

Tip: Allow for variable sensory experiences – if the child has to react to a noise, don't assume that because you heard the noise the child should respond.

Approach Withdrawal. The response of a child to a new situation – whether he is eager to take part or whether he is withdrawing

Tip: Don't assume that reticence by the child is connected to the product being evaluated.

Adaptability. The ease with which a child can change to new things

Tip: Learn to judge how comfortable the child is with any new task. Provide reassurance and extra help when it is indicated.

Evaluators need to be aware of the following three temperamental dimensions that are difficult to control or design for, but, by choosing a good sample of children, can be evened out. There is sometimes a problem with 'teacher-selected' samples in that the teacher (wishing to help) may select persistent, intensive, optimistic children.

Persistence. The length of time a child will give to a difficult task

Intensity. The energy level of a response

Mood. Whether a child is pessimistic or optimistic. Children's motivation may influence the outcome of usability testing procedures (Hanna, Risden and Alexander, 1997).

Other Characteristics of Children

There are some other general characteristics of children that need to be taken into account when carrying out usability evaluations. The following section is an excerpt from work by Markopoulos and Bekker (2003).

Language. Children have developing capacity to verbalise (both vocally and in writing). Many usability-testing methods require the child to either write or verbalise their experiences.

Children may simply not be proficient enough for the method. Nevertheless, to this point there is no evidence that the effectiveness of a usability test in uncovering usability 'bugs' depends on the child's ability to verbalise.

Extroversion. Some children may be more or less talkative. Further they may differ as to how much they are used to speaking to adults and this can affect how likely they are to report usability problems. Evidently, more talkative children will contribute more comments and help identify more problems with a product.

Trustworthiness of self-report. Druin (1999) suggests that children are very honest in their judgements about products, but sometimes the reliability of reported data is questionable (Hanna et al. 1999). For example, children may say they hate the 'bad' character that may in fact be crucial to the success of a product. Alternatively, they may simply name problems to please the evaluator who is looking for them, or they may conceal problems if they think they would offend the software creators.

Monitor progress towards a goal. Children slowly develop the ability to monitor progress towards a goal, to assess the outcomes of a task, and to learn to redirect unsuccessful efforts. The level of this ability will influence how appropriate it is to set tasks to children during a usability test[1].

Knowledge and skills. The ability to understand and carry out instructions for the test, and to interact with more complex products, varies and develops with age, as do the children's interests. Instructions and tasks must therefore be adapted along with the product under test.

The Relationship of the Child to the Product

When we are devising a product for children, we need to consider the relationship of the child to the product, that is, the role that the child is expected to assume when using this product. In a traditional user needs analysis, this stage of enquiry would be 'determining user goals' but often when children use interactive products they do not have a goal, and so we have chosen to avoid the word goal and instead talk about roles. We have identified four roles

The child as a learner

The child as a developing human

The child as a user

The child as a player

Designers and usability professionals can find these roles useful in defining the goals and features of the product as well as in selecting the proper usability evaluation methods to use. By considering these roles, we can identify how children differ from adults in each of these four roles.

[1] Some preliminary evidence reported by Barendregt, Bekker & Speerkstra (2003) suggests that children will report as many problems during free exploration as with when given tasks. They will visit though more screens when given tasks covering a bigger part of a GUI. For evaluations of games, Barendregt et al. suggest not giving tasks unless they are also goals for the game, to avoid that the usability test interferes with the actual playing experience.

Learners

- Children have more to learn
- Children learn more easily
- Children are into learning
- Children' models are incomplete

Developers

- Children age more quickly
- Their brains are still being connected
- Positive rather than negative development
- Physically gaining control rather than losing control
- Used to finding things that they cannot do

Users

- Different motivations, more discretion
- Rarely compelled to use products
- More likely to use products by choice
- Expect more? Magic?

Players

- Children find play very natural
- Play is essential for children
- High levels of imagination

These features make working with children both fun and challenging. Products that we design for them need to take into account these features, and heuristics should ensure that these features are supported in interactive products for children.

AN OVERVIEW OF EVALUATION

This section of the notes attempts to define some terms, and summarises evaluation methods in general, with brief comments on their suitability for evaluation situations involving children. There is more detail of the most relevant methods in subsequent chapters.

Most introductory HCI books include discussion of a range of methods, mainly with adult users primarily in mind, Dix et al (2004) is particularly recommended.

When is evaluation done?

In a product lifecycle (see Figure 1) evaluation is typically presented as the last event in the development process, but evaluation may happen early in the lifecycle as well as at the end. Early evaluations are often called **formative** evaluations as they provide direct input to the design of an artefact during its creation. The thing being evaluated is some form of design specification or prototype. Formative evaluation is usually regarded as being part of the design stage. The term **predictive** evaluation is sometimes used to describe methods used at a very early stage when only specifications are available, rather than prototypes. Alternatively, the term predictive evaluation may be used (as it is in Appendix 1 of this document) to mean any method that doesn't involve users or user representatives directly, i.e. the 'guideline-based ' and 'task-based' methods described below.

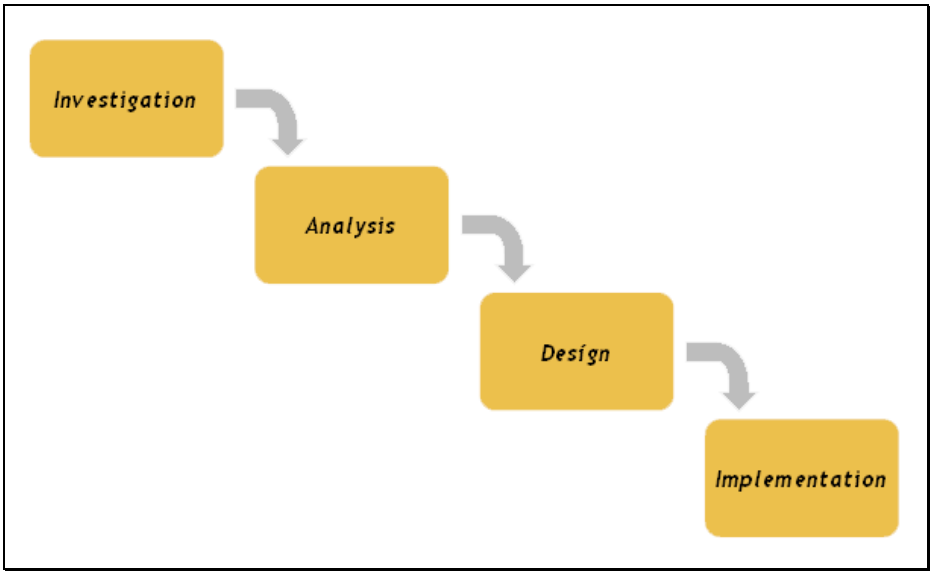

Figure 1. A product lifecycle.

Evaluations done at the end of the lifecycle are called **summative** evaluations; their purpose is not to influence the design of the product – it's too late – but to provide information relevant for the development of the next version, or a related product.

What is being assessed?

Where the intended users of a product are adults, it is most likely that it is **usability** that is being assessed. When the users are children, usability is still important, but it is possible that

the evaluators are also interested in assessing how much **fun** the users have, or how the software supports learning or development.

Although it is widely used in HCI, 'usability' is a difficult word to define. It is generally taken to refer to the ease with which users can complete their tasks, how easy it is for them to learn, and how satisfying the interactions are; this is the definition we use in these notes.

Usability is concerned with the user's attempts to carry out tasks. Young children using interactive products often don't have clear tasks; they are concerned with having fun rather than with trying to achieve something specific. (Note that with educational products, there may appear to be a task, but this task is actually the aim of other stakeholders such as parents and teachers, rather than of the child users themselves.) Fun is not the same thing as user satisfaction, which is task-related.

Both fun and usability can potentially be assessed using the methods discussed below, though not all of them will be appropriate in any particular situation. It is obviously important to decide what it is that is being assessed before deciding on the methods to use.

It is also possible to evaluate other factors, such as accessibility, safety, or usefulness (of, for example, an educational product. Specific techniques for evaluating these are not covered in this tutorial, though many of the methods used are similar to those discussed below, and most of the guidelines we give will still apply.

What methods are available?

Many classifications of methods are possible. Here we divide methods into those based on **guidelines**, on **tasks**, and on **users**.

Guideline-based methods include **expert reviews**, which are done by usability experts who are not members of the design team, and **heuristic evaluations**, which normally are carried out by the design team. These methods do not require the presence of 'real' users. Both methods assume that there is a substantial, and largely agreed, set of interface design guidelines that apply to the particular situation under investigation. Where the users are adults, this is now a reasonable assumption for a wide range of situations, and these are cost effective and useful methods. Where the users are children, however, there is much less established 'theory' about interface design. We are in the early days of Child-Computer Interaction research. Consequently these methods are of less use than they are for adult users. Guidelines for designing 'fun' are particularly difficult to establish.

Task-based methods include predictive **analytical methods** based on task analysis and performance data, and **structured walkthroughs**. These are normally carried out by members of the development team, and again they are cheap since no 'real' users are required. They attempt to find problems with the interaction design by predicting how the users will respond in particular situations. But both methods assume an understanding of the users' tasks, and of the ways in which the users will react in particular situations. As mentioned above, where the users are children, it is often hard to pin down what their tasks are, and it is very difficult to predict their responses to a particular situation. There is less data available to work from, and in any case children seem to be inherently less predictable than adults. Also, these methods are not promising for evaluating fun. So these methods, too, seem inappropriate for evaluation of products for children.

User-based methods are those that involve real people who are representative of the intended users. **Observation** methods involve watching users using a product or prototype, these

include '**usability test**' methods, **'think-aloud'** methods, **co-operative** methods, and **peer-tutoring** methods. These can be successful with children; they are discussed in much more detail in the next section of this document. They can be used in a variety of situations for both formative and summative evaluation, but they are labour-intensive, and hence expensive. A major factor in their success or otherwise is the environment in which they are used; this is discussed below. **Survey** methods involve asking users questions, either in an **interview**, or by **questionnaire**. Sometimes the survey takes place after a user has experienced the product or prototype. They can be used for predictive, formative, or summative evaluation; they are a good method for assessing users' opinions and feelings. Used with care they can work well with children; they are also discussed further below. A variation of questionnaire methods is based on the use of diaries. **User diaries**, used to record critical incidents in real life product use, are also used for summative evaluation. This method has been used with children; more detail is given below.

There are other user-based methods that are less likely to be successful with children. **Focus groups** are groups of potential users who meet with a moderator to discuss formative aspects of product design. **Helpdesk data** can be used for summative evaluation, for example to find common problems. Automated **logging** of users' use of software can be useful in finding patterns of use, common errors, etc.

How to plan an evaluation

There are three stages in selecting an evaluation strategy:

1. What are you trying to achieve?

e.g. to establish what really happens in the real product environment
to compare possible designs
to check whether some pre-defined target has been reached
to check for conformance to standards

2. What standard of evidence is required?

e.g. for an academic publication, may need evidence of a statistically significant difference
for a design decision in industry, less conclusive evidence may be sufficient

3. What do you want to measure?

e.g. time to complete tasks
error frequency
user satisfaction
fun

4. How can this be done?

usually a combination of methods is necessary
taking into account constraints, eg time, money, user availability.

HEURISTIC EVALUATION

We can use heuristics to determine how well a design fits with previously defined standards and guidelines. This sort of evaluation is typically carried out by experts in the domain under investigation (generally usability experts) and it is common to have a list of heuristics together with, in some cases, a space for comment and for a severity rating. It is possible to derive heuristics fro any attribute that is considered to be essential in a system. For products for children, some attributes include fun, usability and pleasing. Others could include educational, accessible and captivation.

Heuristics for Usability (Nielsen 2001)
Visibility of system status – does the child know what is happening?
Match between real and system world – does it match the child's metaphors?
User control and freedom – can the child move around freely?
Consistency and standards – is the interface consistent
Help users find, fix and recover from errors – is it easy to recover from mistakes?
Prevent errors – are there validation and safety measures built in?
Recognition rather than recall – does the child have to remember too much
Flexibility and efficient use – are there short cuts?
Aesthetic / Minimalist design – is it nice to look at?
Help and Documentation – is this child friendly?

Heuristics for Fun (Malone and Lepper, 1987)
Challenge – not too easy, not too hard
Feedback – to reduce uncertainty
Curiosity – video, audio or other stuff
Fantasy – evoke mental images
Choice – should be provided
Control – optimal for user

Heuristics for Pleasure (Preece, Rogers, Sharp (2002))
Satisfying
Enjoyable and fun
Entertaining
Helpful
Motivating
Aesthetically pleasing
Supportive of creativity
Rewarding
Emotionally fulfilling

EVALUATING PRODUCTS BY OBSERVING CHILDREN

Objectives

Observing for the purposes of evaluating a product, can take many forms. It might take place in the laboratory as part of a usability test or it might be in the field as part of an ethnographic study. The data collected might be verbalisations of users or observations made by the observer. Next to the typical considerations one makes for adult users (for which we refer the reader to standard textbooks such as Dumas and Redish, 1993) the involvement of children as testers of products can impact the choices to be made. Further, the social context of the observation is also important. For example, is the child alone? Is there another child present and in what capacity? Is there a teacher or a parent involved? Some of the possible answers to these questions are discussed below. Finally, the peer tutoring evaluation method that is specifically designed for the evaluation of products by children is introduced.

Observational studies in the Field or in the Laboratory

Usability testing in the laboratory is perhaps the most traditional usability evaluation technique, developed by research in the 80's. Usability testing in the laboratory enables the evaluator to perform rigorous and controlled testing; it makes it easier to collect performance data and other sequential data (e.g., video footage, audio recordings, logs) when compared to a field-test. However, developments in supporting technologies such as portable usability labs make it possible to capture such data also on the field.

In the case of adult professionals, usability testing has been criticised because it filters out contextual factors that impact greatly upon the eventual usability of a product, e.g., the interrupt driven nature of work, the physical and social context of work.

Different but similar problems characterise usability testing for children. Children will use software in contexts such as at home or in the classroom, that are physically and socially very different than a usability laboratory. They will interact with products in the context of play or classroom learning that is not the case in the usability laboratory. Bringing kids to the laboratory entails many practical problems. For example, children will need to be escorted to the laboratory, they will need to adjust to unfamiliar, perhaps unfriendly, surroundings, might be distracted by the paraphernalia of a test-laboratory, recording equipment has to be adjusted to their physical characteristics, etc. Table 2, summarises some of the remarks of (Hanna, Risden and Alexander, 1997), pertaining to usability testing laboratory itself. One potential solution is to furnish a special purpose usability testing laboratory. Such a laboratory has been set up in the Technical University of Eindhoven.

Adapt decoration	Colourful posters, appropriate furniture. Avoid childish objects.
Microphones	Small microphones placed close to the children.
Video camera	Avoid pointing camera straight into their faces.
One-way mirror	Children should not face the mirror.

Table 2. Comments of Hanna, Risden and Alexander regarding the set-up of the usability lab.

The use of video-cameras has been a contentious issue both in the context of usability test and in the context of requirements gathering activities during participatory design (Druin 1999). It seems that currently a consensus is emerging among researchers that children can be comfortable in front of video cameras and after a while will forget about them and will proceed with their activities. During field studies they may even be good operators of these cameras for recording useful data. In a recent study, children were caught 'misbehaving' on camera, and became very self-aware. When assured that the evaluator wouldn't tell the teacher, they continued their activities ignoring the camera.

Whether using standard usability lab equipment or a mobile usability evaluation kit, the observation can help study actions of the child (e.g., click-stream, non verbal cues for their emotional state, etc.). This type of observation does not provide access to their opinions and preferences or to their thought processes. For opinions and preferences survey methods should be used (E.g., post task questionnaires or interviews) which are discussed in another section. The think-aloud technique can be used to gain access to the latter. This is discussed further below.

The KidsLab in Eindhoven

At the Eindhoven University of Technology (TU/e) a special purpose laboratory has been set up for conducting product evaluations with children. This laboratory is in most ways similar to standard usability laboratories. The layout is similar to a standard usability laboratory, having a one-way mirror to separate the observation room from the main area. The main area has been furnished with child-sized furniture that can be considered roughly suitable for children up to 7-9 years of age. (At about that age, adult sized furniture is no more a problem). The decoration is colourful and pleasant and curtains have been placed to conceal the one-way mirrors when appropriate. Care has been taken not to go overboard with childish decorations or toys. Figure 2 shows different views of the room.

Figure 2. Different views of the main area of the KidsLab of TU/e. Note the cameras mounted close to the ceiling corners. To the right, the desk is shown with a PC on it. The same desk is seen in the picture to the left, with one camera placed directly above it and pointed to the keyboard. In both pictures the microphones can be discerned on the highest shelves close to the cameras.

With regards to the recommendations by Hanna, Risden and Alexander (1997), shown on table 2, we note the following. Cameras are placed high not facing the child. Two directed microphones have been placed far from the child (each yard lower than a camera), are sufficient to capture the children's voices. The observation room is similar to a standard usability laboratory with mixers for audio and video signals arriving from the cameras and the PC. Cameras can be remotely controlled.

The laboratory is very new and so far only 6-7 studies have been conducted within it, involving approximately 10 children each. Experimenters who have made use of the laboratory appreciate the furniture size, had no complaints about the audio-capture system. Further, the permanently mounted cameras with their remote controlled panning, tilt and zoom remove some of the practical troubleshooting of setting up video cameras in a field situation. Regarding the cameras, it was reported that children ignored the cameras completely even though the researcher made a point of showing them in the beginning of the session[2]. In some occasions the observation room has been used as a sort of waiting room for the parent and siblings of the test-participants. The room mirror allows the parent to observe the child without interfering with the experiment and to be at hand's reach if needed.

The usability evaluations conducted in the KidsLab have so far all being part of research projects of TU/e and of Philips Research. Considering TU/e research into this area, this laboratory has been a valuable infrastructure and a useful investment. Still in the majority of the evaluation studies, field testing in the classroom has been preferred for the reasons mentioned earlier. The laboratory though does provide significant advantages especially when the product tested is not robust (the classroom evaluations concerned off-the-shelf products).

[2] W.Barendregt, personal communication.

Figure 3. The observation room of the KidsLab.

Testing with children in the field

Testing products in the field, i.e., in the intended context of use by children has the advantage of ecological validity (as for adult users), but further puts fewer demands on the children to adapt to new environments. This can be particularly important for children under 9 years of age. Practical reasons also make field-testing attractive. It takes much higher organisation effort to bring 8 children one day to your lab, rather than going to their classroom for a test session. In the latter case, the cooperation of the teacher can be valuable for practical issues like getting informed consent from parents, making time available, e.g., during extra-curricular activities, helping with the selection of the children, finding a location in the school, etc. Further, teachers are useful informants for explaining the researcher's observations and impressions of the children, as they are aware of the children's capabilities, characters and social behaviour, so they can place the observed behaviours in a wider context.

For several products the intended context of use is at home where the parent and not the teacher will provide technical support and explanations and where software is used for fun rather than as a component of the educational curriculum. Arranging a usability test at people's homes can be challenging, so a good compromise is to also test fun products in the school. While the context of use is not exactly the intended one, it is still a familiar and relaxing environment, e.g., during after school activities and one that offers the practical advantages mentioned already.

Usability testing of products intended for classroom education is more challenging. Rode et al (2003), discuss the constraints of designing interactive products for use in the classroom. In their paper they discuss participatory design of technologies with children and they propose four key constraints that must be met by such technologies also in the case of usability testing in the classroom.
Technology must perform:

- In a curriculum-focused context. The usability test should fit within a session and support a lesson plan consistent with the curriculum (e.g., for the UK sessions have to fit 70 minutes for 11 year old children) and they should help children learn content that covers a relevant area of the curriculum).
- In the classroom environment (using the technology available in current classrooms, test sessions should have a high degree of structure, continuity and predictability

regarding the progress of the lesson, the lesson should still go on in case of technical or usability failure)
- Within the administrative and regulatory framework of the school.
- Without introducing 'lesson stoppers' created by the design. Rode et al. (2003) use this term to describe events or materials like games, animations, technology failures, etc., that distract children from curriculum objectives and could derail a lesson.

The evaluator as an observer in a classroom may be interested in several different things. For example, if the aim is to evaluate how a mature product integrates with other classroom activities, the observer will need to take a more active role and talk to children and teachers. If the aim is simply to observe its unencumbered use and low level interaction problems the observer will assume a more passive role.

Children thinking aloud

Think-aloud is a technique originally developed by cognitive scientists to study human problem-solving mental processes. The technique requires people to verbalise their thoughts rather than reflect upon them (Erikson and Simon, 1985). Think-aloud has been touted as 'the' most useful usability testing technique. However, the pure form of verbalisation advocated by cognitive scientists is very different from the one used in actual usability testing (Boren and Ramey, 2002).

For usability testing the user is supposed to provide a running commentary on their actions (concurrent verbalisation). Sometimes, this commentary may be retrospective where a user comments on the video footage of their actions. The former offers more guarantees of capturing actually encountered usability problems and a more faithful account of users'actual thoughts. The latter frees the user from the workload of verbalising while they interact with an unknown product and helps focus on more sever usability problems than concurrent verbalisation.

There is little experience of using think aloud for children (whether for usability evaluation or other purposes). The method has been applied for studying second language comprehension or for studying search strategies in digital libraries. The extra mental workload required for verbalising and the uncomfortable social situation of having a quiet observer telling you to 'keep talking' make it a hard method to apply, particularly for children users. Children are most likely to stop talking when they experience difficulties as they will be busy trying to tackle these and unable to provide a parallel commentary. Further, if they are shy about verbalising their thoughts they become even more so during concurrent think aloud and experience it as an embarrassing situation.

However, Donker and Markopoulos (2002) have shown that children 9-12 are able to provide a running commentary during interaction and indeed that think aloud helps identify more usability problems than a post-task interview or a post-task questionnaire. Rather than a bland request to keep talking the evaluator is required to remind the children to describe the difficulties they experience and sometimes encourage them to go on[3]. In this case, a think-

[3] Donker and Markopoulos (2002) provided 3 questions as prompts to users to think-aloud and asked to provide comments relating to these questions:
- Did you need help solving the task?
- Did the computer make it easy for you to do the task?
- What, if anything, happened that you did not expect or want?

aloud closer to a dialogue as suggested by Boren and Ramey (2000) rather than the monologue that Erikson and Simon (1984) would recommend.

Currently, research is still too sparse to suggest the optimal practice for thinking aloud by children. Pilot testing for the intended target use and social context, will perhaps be the most informative to help the evaluator decide:

- The appropriate complexity and number of tasks for the children
- The appropriate prompting strategy and interaction between adult evaluator and child.
- Whether the child seems unhappy (due to the social or physical environment, the tasks they are asked to do, etc.) and how can this be improved by modifying the above strategies.

One or more children test participants at a time?

The aforementioned problems of think-aloud have lead researchers to hypothesize that usability testing techniques involving more than one child, e.g., co-discovery, where children discover together how to use a product, would provide a more natural context for the child to verbalise. Co-discovery has been introduced for adult users where the ability of adults to co-operate is assumed.

Van Kesterenet al (2003) studied several variations of the setting in which children 6-7 are required to verbalise. Their study suggested that Active Intervention (where the evaluator prompts the children for explanations of what they are doing and to give a commentary on their interaction) seemed the most promising approach. Co-discovery was not as successful because children failed to collaborate. Retrospective verbalisations appeared very useful and promising.

Intervention and interaction with children should be approached carefully. They are very likely to show social desirability bias. To avoid this it is important to state clearly that the aim of the evaluation is to assess the product not the children. However, the usability test should not become a contest of naming as many problems with the products as possible. Further, the evaluator should be careful of not assuming an authoritative role in their interaction with the children but rather become a mate or cooperator. This is very hard to achieve, especially with younger children as both adults and children easily slip into a power relationship favouring the adult. The age of the experimenter may be itself an influence, their dress code and the existing power relationships between adults may all be interpreted by the children. (Hanna, Risden and Alexander, 1987)

What to observe?

As for adult users the evaluator has to make a plan of what events to observe and what utterances to record according to the goals of the evaluation. Contrary to products intended to increase the productivity of adult workers, products for children may combine roles of the child as learner, player, developer together with that of a user. The goals of the usability evaluation should then be set accordingly.

The specific questions were selected in that case to be possible to enable experimental comparison of the effectiveness of think aloud, post task interview and post-task questionnaire, which also used the same questions. In general, more open questions should be used during thing aloud.

Signs of the child (as a player or learner) having fun can be observed when they continue using a product ignoring requests to change it, when they become oblivious to their environment. Signs of the child not having fun could be sighs, smiles, fidgeting, etc. Even an absence of any positive emotions should be noted, either as a problem with the product or the evaluation set-up. More on observing signs of the child having fun can be found in another part of this tutorial describing the 'fun toolbox'. It is however not straight forward to distinguish between problems/frustrations caused by lack of usability or fun. Barendregt and Bekker (2003) describe a scheme for classifying observations relating to lack of fun and usability.

The focus of the observer changes when we consider the child as a developer or a learner. Price et al.(2003) postulate the importance of playful learning. During their evaluation of an experimental product they search for signs. In addition to fun, they search to observe aspects of:

- Exploration through interaction
- Engagement
- Reflection
- Imagination, creativity and thinking at different levels of abstraction
- Collaboration

To assess the educational value of the product, the goals of the evaluation are influenced by the pedagogy behind the product, e.g., whether it is supports constructivist learning or reinforcement learning and the educational aims it supports. So further to testing whether the product helps meet educational goals signs of the child exhibiting desirable behaviour should be observed.

Peer-Tutoring

This is an abridged version of the results that have been published elsewhere. The full reference of the paper is Höysniemi, J., Hämäläinen, P., and Turkki, L. (2003). Using Peer Tutoring in Evaluating the Usability of a Physically Interactive Computer Game with Children. Interacting with Computers, Vol. 15/2, May 2003: Special Issue: on Interaction design and children. pp. 203-225.

Introduction

Peer tutoring is a usability evaluation method where *children teach other children* how to use the product that is being evaluated in a familiar social setting (Höysniemi et al., 2002, 2003). The method is based on a well-known approach in children's education (Topping, 1988, Garvey, 1986, Goodblad & Hirst, 1989, Gaustad, 1993, Rogoff, 1990) but it was first developed and applied in usability evaluation purposes of the physically and vocally interactive computer game for 4 to 9 year old children (Höysniemi et al., 2003). In addition to peer interaction studies the approach has benefited a lot from traditional usability research (Nielsen, 1993, Dumas & Redish, 1993), and especially the studies and guidelines (Hanna et al., 1999, 1997, Druin et al., 1999, Cassell & Ryokai, 2001) made for the design of children's computer products. The peer tutoring is based on observing whether children are able and willing to teach other children to use the product that is being evaluated. In addition, the method provides information on how children use the system and how they communicate about the product.

Definition

Peer tutoring is one type of peer collaboration. Damon and Phelps (1989a) define peer tutoring as an approach in which one child instructs another child in material in which the first child is an expert and the second is a novice. The peer tutoring method differs from the co-discovery learning and co-participation methods in that it is not based on the idea that two participants work collaboratively on a given task exploring a system together but that the task of a tutor is teaching and the task of a tutee is to act out according to the tutor's instructions. In the study of Höysniemi et al. (2003) both the child who is the teacher (tutor) and the child to be taught (tutee) are not very far removed from competence due to the tutor's short training period.

Process

Two modifications of the method have been applied depending on the number of tutors. In *two-on-one tutoring* two children teach one tutee and *each one teach one method,* where each tutee acts as a tutor for the next child, and thus transfers his or her experiences and knowledge about the game to the next child. The peer tutoring process usually consists of four phases: (1) the introduction of the test setup and of the tutor-tutee roles to evaluators, (2) the training of the tutors, (3) tutors teaching the tutees, and (4) final interview.

During the tutoring session Höysniemi et al. (2003) suggest that the researchers do not teach or instruct the tutees but ask questions from the tutors if the teaching situation requires adult intervention. The *question-asking protocol* combines both interviewing the tutor and provides help for the tutor in the teaching situation. The question-asking protocol is simple and effective. Two categories of questions are used: (1) questions that help tutors to teach a tutee and carry out given tasks like "could you explain Paul how to send email?", and (2) comment related questions (example dialogue: first tutor says: "it is very tricky", an adult asks: "what is tricky?", and then tutor replies: "it's very difficult to find a button where to send the email"). When answering a question, the tutor provides product related information based on the tutor's own experiences and observations using language similar to the tutee's.

Benefits of the method

The notion of the usability of interactive systems depends on the social context in which the interactive systems are used. An unfamiliar environment like a usability laboratory and observational technology may accentuate children's awareness of being judged and observed despite reassurances that it is the software that is being tested. In the peer tutoring approach testing can be carried out in the natural environment of the children – homes, schools and day care centers with age mates they already know and choose to work with.

In a child-adult relationship, the differences in the levels of knowledge and authority affect the nature of the discourse between the child and the adult (Damon and Phelps, 1989a, p. 138). Thus, the challenge in an observational situation is to open up the communication and encourage children to be verbal and diminish the authority. Also shyness, the fear of giving wrong answers and children's need to please adults can affect the test results. Peer collaboration provides tools for communication and equality. "Children have certain advantages over adults in teaching peers. They may more easily understand the tutee's problems because they are cognitively closer" (Gaustad, 1993). Rogoff (1990, p. 172) also implies that children are likely to treat the situation differently if they are in charge of it rather than being given a task by adults, which is the case in traditional usability testing.

Peer tutoring provides us with information about the learnability and teachablity of the system and what kind of instructions children use when teaching one another. In practice this means that tutor takes an active role as an instructor in a usability testing situation. The peer tutoring situation is not, however, adult-free since researchers are still in the same space to help and guide the tutor-tutee interaction but only if necessary. The researcher-child communication requires careful planning before the testing. Especially younger children need help in teaching the other child, which requires some adult intervention. Since the test situation is not completely adult-free, it is important to make sure that the researchers behave informally and make space for the children to interact with each other. Since children are more active in the testing situation, researchers can stay more remote and adult intervention is diminished.

The communication in traditional usability testing is likely to be asymmetrical between adults and children. The peer tutoring approach encourages children to verbalize their thoughts naturally and spontaneously. One benefit of peer tutoring is that peers speak in a more similar manner than do adults and children which also holds true to other collaborative usability methods as well. Also, children are more relaxed to communicate with adults when there is a peer in the test space. The tutor-tutee communication is highly valuable when analysing how well children have learned the required skills needed to use the product, how they perceive the interaction and how much and what type of instructions are suitable for children.

Drawbacks

One drawback of peer tutoring is that tutees, often labeled as less capable than tutors, tend to resist being tutored by their age mates (Gaustad, 1993). This problem can be reduced by not using tutors that are much more skillful than the tutees. The communication between tutor and tutee can become unbalanced meaning that tutors speak much more than their tutees. As with many other qualitative evaluation methods the peer tutoring approach also requires a lot of work in organizing the test sessions and analyzing the video material.

Tutors tend to take their role seriously and want to teach their tutees correctly. Thus, the research challenge is not only to find out what things children can teach to each other but also what they leave unsaid and why. Children usually first explain the things that they like and find funny leaving out the complex and boring features of the product. Therefore researchers must be careful to get children to verbalize their thoughts about design solutions that they do not fully understand or like. One possible approach might be to use a visual walktrough or to combine group play and interviewing to the test setup.

Also, as Van Kesteren et al. (2003) mention that “the researcher should be aware that if the tutor forgets how the task works (and they sometimes do) the tutor-tutee situation changes somewhat in that the children will work together, as is the case during Co-Discovery”. This requires researcher to plan their actions beforehand and choose how to guide the situation towards the peer tutoring model.

Depending on the product to be evaluated and the test setup, the tutors can take over the task completely and not let their tutee complete the task. This drawback can be dealt with by explaining to the tutors that they can only explain and show but not carry out the tasks themselves while teaching. Also, a careful test space design helps to tackle the taking-over behaviour. One option is to locate the children so that it is not too easy for a tutor to start operating the software.

The problem with the question-asking protocol is that the tutors tend to direct the answers to the interactor rather than the tutee. The tutee, however, listens to the tutor's answers carefully and uses the information in these answers to understand the functionality of the product better. Additionally, tutees pass the knowledge obtained from the tutor's answers to their own teaching and answers as seen in each-one-teach-one approach.

The adult intervention in the peer collaboration approach is problematic. It is difficult to be assertively passive while guiding the interaction. It is also hard not to get carried away when the children are very enthusiastic or to cheer them up when they are bored. The interactor in the test situation should not interfere with or lead the collaboration between the children any more than is actually required.

Variations of test-protocol during usability testing

This section discusses different methods for user based (empirical) usability testing. Different methods differ mostly in the way the test participants engage with the product (e.g., having set tasks to execute, or free exploration of the product), they way they interact with the experimenter (e.g., no interaction, dialogue, question answering), whether they are alone or with another tester and finally in the manner that data is collected.

How to collect data

During user observation a researcher observes children as they use an interactive product, and take notes on the activity that takes place. Observation may be either direct, where the adult is actually present during the evaluation, or indirect, where the activities are viewed by some other means such as through use of a video recorder. Usually user observation does not necessarily require tasks to be given to participants.

Another means to obtain usability data is to record and analyse interaction logs. For example, the click stream of the user or their eye movements might be analysed to obtain usability information. Also, performance measurements can be taken to obtain more objective measures for comparing different designs and for assessing the usability of a product.

Usability test can involve providing subjects with tasks and observing them while they perform given tasks. This is the typical practice for adult users. For children, it may be more appropriate to allow free exploration, especially when a critical component of the interaction that has to be evaluated is that of fun. For example, Barendregt et al (2003) found that for evaluating computer games for children it is not advisable to set them test-tasks, as these will compete with the internal goals of the game and will not allow the children to have fun with the game during the evaluation. session.

Observing the children when they experience problems with an interactive product or when they are getting frustrated with it, does not always help identify the causes for these problems. This limitation is particularly relevant in the context of usability testing in order to be able to suggest improvements for the interaction design of a product. For this reason, usability testing very often requires to obtain self-report data from test-subjects in the form of an interview or even verbalisation of their thoughts. The latter can be obtained using a range of techniques that are described briefly below.

Thinking aloud

During the course of a usability test, the test users are asked to provide a 'running commentary' their thoughts, feelings, and opinions while interacting with the system. The evaluator should not interfere. The purpose of the thinking aloud is to get information about children's mental processes and thoughts. According to Dumas and Redish (1993), a proper use of the method requires instructing children prior to the interaction, having them practice with thinking aloud and prompting them to do so during the testing. Usually thinking aloud protocol is combined with user observation and is usually done with pre-set tasks for the user.

Active intervention

During a usability test participants are required to verbalize their thoughts and the testers prompt them by asking direct questions about the product. This is supposedly a more natural way than the thinking-aloud method in letting participants verbalize their thoughts.

Retrospection

The retrospection method consists of performing a usability testing session while recording it on video and then watching parts of the recorded video together with children whilst having subjects respond to questions about their interaction. The child's comments while reviewing the tape are thought to be more extensive than the comments made during the test activities. The retrospection method also gives researchers the opportunity to stop the tape and ask questions about certain actions or comments on a tape.

Co-discovery

In the co-discovery method, pairs of test users perform tasks with a system, which they discover how to use, together cooperatively while being observed. They are encouraged to explain what they are thinking while working on the tasks. This technique should make it more natural for the test users to verbalize their thoughts during the test.

Peer tutoring

Peer tutoring is a usability evaluation method designed for evaluating products together with children. The peer tutoring is based on the idea that children teach other children how to use the product that is evaluated (Höysniemi et al., 2003). In peer tutoring each child participates in two sessions; in the first session a child practices the tasks and becomes experienced in using the product, and in the second session he/she instructs other child how to carry out the tasks.

Comparisons of the methods

Table 3 below compares different techniques for usability testing, particularly focusing on different protocols for the execution of the test and the collection of the data. Note that the following table only contains the methods that have been reported in literature to be used in evaluating products with children. The table below is an adaptation and extension of the matrix of Van Kersteren et al. (2003).

Method for obtaining usability data	Skills children should have or be able to do	Benefits	Drawbacks	Additional information
User observation	No special requirements .	Can be conducted in a child's natural environment. Suitable for young children. The most natural form of field testing.	Does not necessarily focus on the most interesting features of the product. The data collected might not give any measurable data.	Not necessarily task-based. Usually within a field study. Can be applied as early as concept design phase.
Performance measurement	Use the product individually. Understand and execute task descriptions.	Can provide objective data. One of the few quantitative evaluation methods.	More attuned to adult users and productivity oriented products. Children might feel that they are being judged by their performance.	Task based. Needs to be combined with another method for getting satisfaction/fun measures. Better for later stages of product development. Experimental design very important.
Post session or post task interview	No special requirements from the method.	Avoids distracting children during the interaction. For younger children recall after the session may be difficult, so asking questions soon after a task may be beneficial.	Requires interviewing skills from the experimenter. Requires careful design of questions.	May or may not be task based. Asking questions after an interaction session where the experimenter acts as an observer is the most usual practice. Can be used at all phases of design. For younger children the test session can be divided into smaller sections followed by an interview.
Thinking aloud	Use the product individually. Be able to verbalise thoughts while at the same time interact.	Seems to be very efficient in uncovering usability problems with respect to the number of children involved. Not influenced by other child's presence in the test space.	Thinking aloud needs practicing before the test Think aloud is unnatural for the child. The language children use might be simpler than their mental processes. Analysing children's verbalizations require skills.	It is usually task-based. Think-aloud can be performed with different degrees of interactivity by the experimenter. For children, it is recommended that a more dialogical style of interaction with the tester is adopted. Better for later stages of product

			High cognitive load.	development.
Active intervention	Answer questions about preceding plan, actions and evaluations. Perform tasks individually.	Allows researcher to ask about activities when they happen. Children verbalize their thoughts easily. Children might feel that their opinions are important.	Based on predetermined set of questions. Researcher may bias the child. Not very flexible, i.e. it requires skills to be able to prepare for the unexpected activities.	Is usually task-based. Better for later stages of product development.
Retrospection	Answer questions about recalled plans, actions and evaluations. Recall what is done before and be patient to watch the session again. Perform tasks individually.	May offer insight to higher level problems than think-aloud. Offers insight into thoughts of the tester without incurring the extra cognitive workload during interaction.	User does not necessarily remember what he/she was thinking while performing tasks.	May or may not be task based. Can be used at different phases of a design project.
Co-discovery	Express themselves to other children. Perform tasks in co-operation.	Might give more verbal comments than thinking aloud. Children are more relaxed when there is a peer in the test space.	Children might be competitive and they might not co-operate. Children can forget the tasks and start exploring the product collaboratively. Might be based on negotiation rather than verbalization of mental processes.	Task-based, Collaborative. Better for later stages of product development. Requires 2 children per session.
Peer tutoring	Teach their understanding to another child. Teach by explaining and not by taking over the tasks. Perform tasks individually with the help of a tutor.	Children are more relaxed when there is a peer in the test space. Children have an active role. Outgoing tutors verbalize their thoughts easily and spontaneously. Can be used to assess learnability and teachability.	Tutors tend to take over. Children's mutual relationship affect the outcome of the test. Tutors need to be "trained". Needs planning logistics for groups of children at a time.	Task-based, collaborative. Based on role play. Works best when a tutor and a tutee have a friend-like relationship. Works best with children that are in school. Requires at least 2 children per session. Testing requires either high-tech prototypes or fully functional products.

Table 3. Comparison of the methods

SURVEY METHODS

A survey method is one that relies on a question – answer approach. The researcher or evaluator formulates a question and the respondent answers. The important steps in the question answer process are

(1) Understanding the question
(2) Retrieving relevant information from memory and 'computing' an answer
(3) Formatting the answer (choosing the appropriate response category – as in a Likert scale)
(4) Evaluation of the answer (may result in editing due to social desirability or peer pressure)
(5) Communicating the final answer

With child respondents, there is an increased chance for error in any of the five stages outlined above.

It is unwise to attempt to survey children that are four or less as they have limited language and thought processes. Children aged 4 – 7 (Intuitive thought stage) can be surveyed, but interviews are better than questionnaires, particularly in the lower ages. Children of this age are very literal, so questions like 'have you used a computer like this before' will be inappropriate as they may interpret it to mean a computer of the same colour, shape, make and size when all you wanted to know was had they used a laptop computer before! This age of child is very suggestible, they will want to please the surveyor and will be reluctant to express their own thoughts or feelings. Their attention span is very short and they easily lose interest. They are more likely to use 'satisfying' responses and approaches when they are not very interested in the activity. (Vaillancourt, 1973)

Children aged 8 – 11 can be surveyed more successfully and can complete simple questionnaires, however, they also tend to be very literal and cannot easily understand negatively constructed questions. Whereas practice with adults favours interspersing positive and negative statements when determining attitudes, this is not advisable with children. Short concentration spans or boredom and may also result in unanswered questions or satisficing answers to questions.

Children with more developed language skills produce better data and research has indicated that low reading ability correlates with the number of unanswered questions. Interestingly, boys are more likely to leave questions unanswered than girls, and the proportion of unanswered questions decreases with age.

Planning a survey

Whether planning an interview or a questionnaire, there are some similar issues to consider, as well as some that are specific to one method or the other.

What is the survey for?

It may be to identify usability problems or it may be to elicit the children's preferences about and / or attitudes to one or more products

When will it be done?

It can be helpful to carry out a survey before children use a product as well as after it has been used.

How will it be administered?

Will it be an interview or a questionnaire? If an interview, will it be one on one, or in a group? If a questionnaire, will the children do the questionnaire on their own or in a group, will they have help with the words?

What will be collected?

Decisions about how the data is recorded need to be made – will audio or video recording be used, how will children be identified / anonymised? How will data be safely stored?

Designing Questions (and Answers)

The key to a good interview or questionnaire is to design the questions VERY carefully. This is the case whether adult or child users are involved, but with child users it assumes greater importance due to the child having a different understanding of language than the adult.

The language that is used needs to be the language used by the children. Practical ways of ensuring that this is the case include carrying out a pilot survey, asking a class teacher, or researching language development. Within a group of children of a similar age there will be significant differences in their individual language skills. For a survey to have validity, it needs to have meaning for the least articulate members of the group.

It is often difficult to get children to comment effectively on closely related things. Depending on their age, some children may be unable to differentiate between constructs such as 'like' and 'find easy'. The younger the children are, the more difficult they find it to state preferences for products. Ideas to help children do this can be found in the Fun Toolkit (Appendix 3). Young children also exhibit a strong acquiescence response bias; that is they tend to say yes irrespective of what the question is! This can be in part overcome by asking them 'what do you feel about' rather than questions that those that result in a 'yes / no' response.

It can help children in the answering phase if visual stimuli or response cards are used. We have used visual Likert scales to effect with children as young as five.

Interviews

An interview may be one on one or may be more of a focus group where there is a group of children and one or more interviewers. Breakwell (2000) identifies a range of hazards associated with interviewing children. He states that difficult people to interview include young children who are unwilling to assert themselves or to contradict the adult. Interestingly he notes that teenager children behave in an opposite way!

When children are being interviewed, they have different priorities than adults – they may not realise that in an interview the adult asks and they answer, they too might want to ask! Children are likely to hesitate and in these instances the interviewer should avoid jumping in.

Early interviews can be used to inform the design of later interviews. Interviews need to be carefully planned and the room or space that is being used should ideally be quiet and free from distractions whilst also being safe for both the interviewer and the children. It is useful

to record interviews as well as make notes during the session. Note taking while the child is talking needs to be done very carefully. The child needs to feel comfortable and he or she may well ask what you are writing down.

Questionnaires

The big problem with questionnaires and children is that the children have to be able to write and read; additionally, as the language has to be simplified to make it understandable, it can be difficult to make the questions clear. It is essential to pilot questionnaires before they are used as they can cause considerable stress if presented inappropriately to children. Class teachers can be helpful in looking over questionnaires to see if they are age-appropriate. In our experience, we believe that questionnaires for children should

Be attractive to look at
Have only 5 – 15 questions on them
Have simple questions at the start to make it easy for the children to get started
Be presented on one side and one sheet of paper only
Be printed in a child friendly font and font size

When questionnaires are administered to a group of children in one place, there will be some collusion going on. This is a disadvantage that has to be outweighed against the advantage of all the children filling in their views at the same time and the time that is saved by doing it this way.

There will be some spoiled questionnaires. It is likely that you will be unable to work out what all the words mean as children experiment with their spellings. Some children will choose to not answer all the questions; this may be for any number of reasons. Some of these reasons are identified here;

they are too tired,
they cannot read the question,
they cannot understand the question,
they don't know the answer,
they don't know how to write their answer,
they are bored of answering the questions,

If at all possible, the child should be helped to answer all questions as their comments may be very enlightening. One method is to ask children as they hand in their questionnaires about any missing answers / comments.

Measuring User Satisfaction with Child Users

Quite often, the purpose of a survey is to determine user satisfaction. Satisfaction measuring with children is particularly problematic when computer interfaces are being evaluated. The computer has a 'bewitching' effect on the child; Janet Read's work on developing metrics for satisfaction has demonstrated this, with children giving exceptionally high ratings for computer-based activities. Difficulties also arise when children are asked to rate products for 'enjoyability' or for 'fun'. We have observed children changing their minds when it appears that one product is getting a 'bad' score over a number of questions. We have also observed children deliberately making things fair over a series of ranking questions.

As has been described above, young children find it difficult to differentiate between similar but different constructs. For example, questions like 'Which was most fun?' and 'Which worked the best?' are not as obviously different to a child as they are to us.

To assist children to articulate satisfaction with products, we have developed a fun toolkit that can be found in Appendix 3. The key 'survey instruments' in this toolkit are a Likert style smileyometer, a repertory grid style 'fun sorter' and an 'again-again' table.

The following grid shows which tools can be used in which situations;

	Comparing two or more alternative products or features	What does it show
Before seeing the single product/ trying the single feature	Smileyometer or Funometer NOTE: Only show one at a time – this will ensure that the effect of playfair is eliminated	The child's expectation of the product / feature
After seeing the single product	Smileyometer or Funometer	The child's experience of the product / feature
After seeing all products	Fun Sorter or Again-Again table The again-again table can be useful for multiple products; The fun sorter is better suited to multiple features. The fun – sorter can be used for multiple products with multiple features but it in this case, it should be presented one at a time (and shuffled in order across a group) to ensure that the effect of playfair is eliminated	The comparative experience of the child across the products / features. The fun sorter returns a ranking for one or more products or features. The again-again table returns a score based on the desirability of the child to use the product / feature again
A week or so later	Remembrance tool – where a large (>4) number of products or features have been looked at, this tool is useful to determine which was least liked / preferred.	Based on the Pollyanna principle – this results in a list of the features / products that the child remembered over time. These are likely to be those that the child most enjoyed

Table 4

USING DIARIES

An alternative way of recording user interactions with a product is the use of diaries. Diaries can be written journals, some times they can be audio recordings (on memo-recorder devices or voicemail) and they can be enhanced with still pictures. Diaries afford higher ecological validity of recorded information; i.e., they record interaction incidents in realistic situations and context over prolonged periods of time. Especially for the domestic use of technology or the use of mobile devices, diaries seem very promising.

Diaries depend on the test-subjects to be willing, disciplined enough and able to record required information about their use of products. This makes it a particularly hard technique to apply for children users.

In a recent study of cross generational communication, children were equipped with specially designed diaries. These were designed to look playful and fun rather than 'scientific' forms to be filled in. Prompts were provided in the margins to show the type of information sought, rather than providing children with a long and tiring list of potential questions. Children were also equipped with a small digital camera in the shape of pen, with which they would capture images relevant to the study (in that case photographing things they wanted to show to their grandparents).

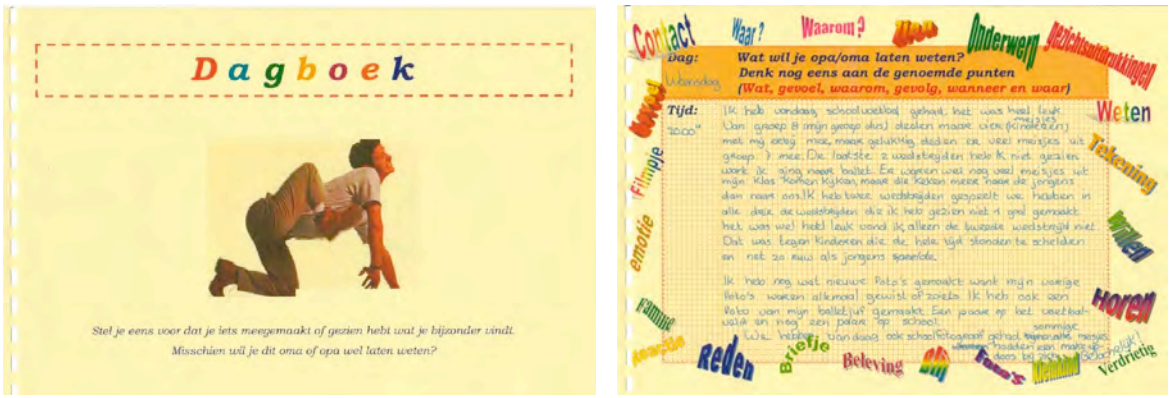

Figure 4. Design of diary for children: To the left the cover page of a diary meant to pre-dispose playfully the child and to avoid stereotypes of the 'scientific' enquiry. To the right a page of the diary completed by a child aged 11.

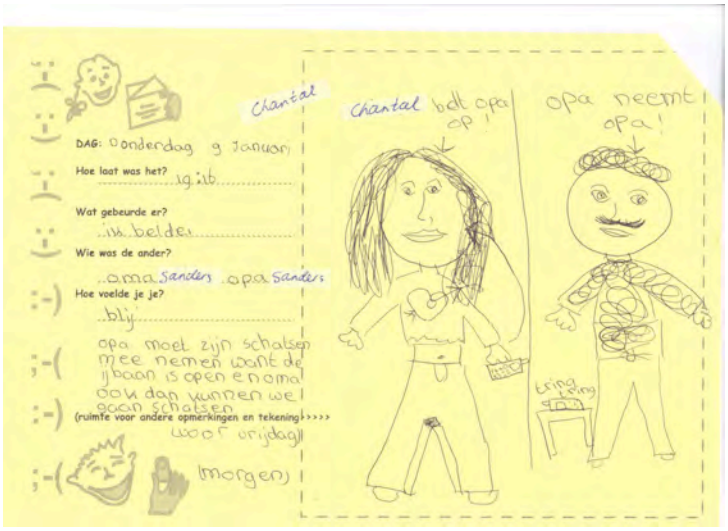

Figure 5. Diary page completed by a child aged 7. In this case, only a few questions were printed on the paper. The child made a drawing of her communication with her grandfather. Diary entries by children were very informative in this study but again very sparse in information content compared to those made by adults.

We still have very small experience of using diaries with children. We found though that even when they are well able to write, children are not as good at expressing thoughts, feelings and preferences, so the evaluator should better ask them to record simple facts and observations. The still pictures they took when given a digital camera were not always very informative. However, children seemed able and willing to carry out the study and useful insights were thus obtained.

In conclusion, the diary method is advisable when ecological validity of the evaluation is of importance. It should be designed to be little taxing to the child and it should be combined with other methods such as logging and interviews for getting an insight into the actual usability problems encountered (van Vugt and Markopoulos, 2003).

ETHICAL CONSIDERATIONS WHEN EVALUATING WITH CHILDREN

Ethics can be defined as "the study of proper action". Ethical decisions are often subjective, and even experienced investigators are likely to disagree about whether a proposed action is ethical or not. If in doubt, err on the side of caution!

Organizations such as ACM and APA provide ethical codes (ACM, 2003, APA, 2002), but there is a shortage of examples how these ethical principles can be applied in practice (Burmeister, 2001, Molich et al., 2001).

There are three main areas of concern where child evaluators are concerned – **safety**, **consent**, and **privacy**. Note that it is not just the children, but also the adults involved who need protection. This protection comes from proper consideration of the ethical issues in a proposed evaluation, in advance of the work.

A note on terminology: an 'evaluator' is one of the children helping with an evaluation, an 'experimenter' is an adult who is working directly with the evaluators, the 'investigator' is the person who designs the experiment, and the 'client' is the person or organisation who wants to know the results. The client and experimenter have a duty to behave ethically, but it is the investigator who is responsible (morally, and probably legally as well) for ensuring that the work is carried out ethically. (In a small experiment, the investigator and experimenter – and even the client - may be the same person, but it is useful to consciously separate the roles.)

In most countries there are laws regulating some of these aspects, for instance Health and Safety, Child Protection, Human Rights, and Data Protection legislation. Of course these laws vary from country to country so it is difficult to be specific about them here. The general trend in all of these areas is that laws change frequently, and are becoming more protective of children. In general, an evaluation experiment that is ethical is likely also to be legal; the reverse isn't necessarily true.

Safety

In carrying out the investigator's experiment, the evaluators should not be at any more risk (either genuine risk or risk perceived by the child) of any kind, compared with their normal activities. This might mean physical risk (e.g. using equipment, travelling to an evaluation lab), or psychological risk (e.g. being stressed by a task that is too hard, or a time limit). These risks can be reduced by careful consideration of the logistics of the evaluation, and by pilot-testing of the intended evaluation tasks. Another safeguard here is ensuring that the children know that they can drop out of the experiment at any time without penalty; see 'consent', below. Also, an evaluation session should be stopped if a child participant is becoming distressed.

A critically important type of safety issue is the risk of child abuse. If an experimenter is working closely with a child, this provides an opportunity for the experimenter to molest or

abuse the child, or for the child to accuse the experimenter of inappropriate behaviour. Vetting of the experimenters in advance is not sufficient to prevent the first of these risks, and does nothing to reduce the second. A better approach is to design the experiment so that an experimenter and a child are never left alone. A second adult in the room can act as a deterrent, and if necessary, as a witness. This might increase the costs, or even reduce the effectiveness of the experiment in some cases, but it is a price worth paying.

Consent

Ideally, participants in an experiment should give their consent. They should be participating voluntarily, and with an understanding of the purpose of the experiment and their role in it. This is called 'informed consent'. Even with adult participants this level of understanding of the situation is difficult to achieve. The purpose might be very complex or difficult to explain. It may not be possible to explain exactly what data is being collected and why, without compromising the data (see 'Deception' below).

Some philosophers argue that even informed consent is insufficient to allow a person to make a valid judgment on whether to participate in an experiment; they say that 'educated consent' is required. This means that the person would need to be in a position to judge whether the experiment has been designed appropriately, and how the results will contribute to knowledge. This level of understanding is probably impossible in most practical situations.

When the evaluators are children, things become more difficult. Young children may not be capable of understanding the purpose of the evaluation experiment, even at a simple level. It is unlikely that they will be able to give informed consent, let alone educated consent. Legally the child's parent or guardian is able to give consent on their behalf. (The age up to which this applies will vary between countries; in some countries, and some situations, the child's schoolteacher will be able to give consent.)

But just asking the parent/guardian/teacher instead doesn't absolve the investigator of responsibility to consult the child too, as far as possible. An appropriate plan in most situations would be to seek at least informed consent from a responsible adult, and also to explain to the child at an appropriate level what is going on. The child has a right to decline to take part (even if a responsible adult has given consent); his/her wishes must be respected. If a child wishes to drop out of the evaluation session before it is completed, he/she must be allowed to do so.

Deception

In some evaluation situations, it isn't practical to explain in advance exactly what the purpose of the test is. For example, the investigator might be interested in children's reactions to unexpected error messages; explaining this in advance would make the data worthless. To do this test, it is necessary to deceive the evaluators; or to put this another way, to lie to children. Some people believe that this is always unacceptable when working with children; others believe that no real harm is done if deception experiments are carried out carefully. An example of an experimental design where some deception is involved is the Wizard of Oz technique, discussed further on chapter "Evaluation methods".

A policy that should be defensible in most situations is that a deception is justifiable when all of the following conditions are met:

(1) the data to be obtained is genuinely valuable

(2) there is no way to collect the data without deception

(3) the real purpose of the experiment, and the nature of and reason for the deception, is explained to the responsible adult who is giving consent

(4) the real purpose of the experiment is explained (as far as possible) to the children as soon as practicable after the experiment

(5) the evaluators are given the chance to withdraw retrospectively. It is possible that, had they known the true purpose, they would not have consented to take part; consequently they should be allowed to say that they don't want data collected from them to be used.

Privacy

The evaluators have a right to privacy. No data collected in the course of an evaluation should be published in any way that allows the identification of the evaluator, without full consent. Even with consent, the investigator needs a good reason to breach the privacy. Note that 'data' here might include numerical results, verbal comments, photographs, or audio or video recordings.

One way to guarantee privacy is not to collect names or any other data that identifies an individual, but this isn't always possible. The next best solution is to keep the evaluation data separate from the identification data, using a code to link the two, so that the evaluation data is anonymised.

General Data Protection principles should also be adhered to, where there is any possibility of individuals being identified. Only personal data that is necessary for the evaluation should be collected. The data should only be used for the purpose for which it was collected, and should be destroyed within a reasonable timescale afterwards (five years is often a reasonable timescale).

Getting Advice and Permission

A problem for the investigator designing an evaluation experiment is that they may be too close to the situation to see all of the ethical problems it raises. It is always useful to get the views of someone else. In most large organisations involved with evaluation and other research involving children, there will be an Ethics Committee whose role it is to advise on good ethical practice, and to refuse permission for procedures that it considers unethical. This can be a useful safeguard for the investigator, but it does not remove the investigator's responsibility!

APPENDIX 1: USABILITY TESTING WITH CHILDREN

Introduction

Evaluation methods that have worked for adults will not necessarily work for children:

- Predictive evaluation. To this point sets of heuristics or models of fun/learning have not been sufficiently developed to support the predictive evaluation of children software. A first collection of guidelines for the evaluation of websites is presented by Gilutz (2002). Inspection methods like heuristic evaluation and cognitive walkthrough may apply directly to evaluate user interfaces, when considering the child as a user. They need to be extended to address the child as a player or as a learner. A checklist for the evaluation of educational software is described by Wartella et al (2000).
- Usability Testing is a tricky procedure that is hard to execute with children. Adaptations and techniques for usability testing are discussed more below. However, higher validity of results and more child-friendly procedures may be obtained by conducting field tests, i.e., evaluate in the daily environment of the child. While quantifiable data may be harder to collect, more dense data may result concerning the appropriateness of a system for the child tester.
- If possible, a quick and dirty usability test can provide very valuable information at lower cost than the other methods. Simply getting a child to use your system can be valuable at all stages of the design.

Guidelines for Usability Testing with Children

Great problem – children so unpredictable it will go wrong.
"*The things we collect statistics about are primarily those things that are easiest to identify and count or measure - which may have little or no connection with those factors of greatest importance."* (Norman 1994) in 'Things that Make Us Smart'.

Usability tests can be carried out which measure the effectiveness, efficiency, and user satisfaction of a system (ISO 9241). The metrics that may be used for this include percentages of mistakes corrected successfully, time taken to complete tasks, rating scales for satisfaction, and time spent correcting mistakes. Measures of effectiveness, efficiency, and satisfaction may require observational and experimental techniques that will typically involve the user.

Work by (Hanna, Risden et al. 1997) has produced a set of guidelines for usability testing with children; these incorporate general advice on the operation of usability studies and also give some advice on measuring the engagement of the children by observing them at the computer. Later work described a range of techniques for usability research with children (Hanna, Risden et al. 1999). These techniques included iterative laboratory tests and longitudinal tests that incorporated questionnaires that were given to the children to complete. The following table summarises some practical considerations for setting up a usability test with children.

Problem caused by Involvement of Child	Possible solution
Planning the Test	
The test may be too tiring or demanding for the children	Choose number and complexity of tasks to fit targeted children. Pilot test to find appropriate tasks. Schedule children for an hour at a time. Preschoolers will last an average of 30 minutes (Hanna et al 1997). Switch the order of tests around between children to avoid fatigue effects (Hanna et al 1997).
Children may not understand test instructions	Adapt instructions to knowledge and skills of children. Consult an expert, e.g., a teacher for the relevant age-group. Pilot test.
Select carefully children to participate	Screen or pre-test for different familiarity with computers, the input devices used, reading skills, the specific application / product domain (e.g., internet, games, etc.). Do not exclude 'disruptive' children from the test.
Test environment is made for adults	Adapt to their softer voices (Hanna et al 1997). Be moderate in adaptations of testing environment for children. Avoid distracting elements.
Getting Informed consent	Parents may need to sign consent forms. This requires extra planning if the testing will be done in the classroom (so that the parents return the form prior to the test).
If testing at a classroom children may discuss pre-test questionnaires	Organise group sessions for pre-test questionnaires even if you do usability tests individually.
Children may be excessively influenced by the instructions of adults facilitating a usability test	Agree and pilot test an intervention scheme: e.g., when to help the child, how to obtain verbal reports from them during or after the test. Even children of 9 can provide a running commentary, similar to 'think aloud'. Redirect questions from children seeking confirmation of their actions (Hanna et al 1997).
Younger children may not be able to follow a specific prescribed task sequence	If possible allow free exploration. Do not ask them if they want to do tasks, they might say no (Hanna et al 1997).
Determine goals of the evaluation	
The role envisaged for the child (user, developer, learner, player) sets different requirements for the system	Evaluating for the player requires assessing the fun experience, e.g., in terms of engagement or pleasure. Evaluating for the child as developer can inform the usability test with respect to specific skills or behaviours that should be enhanced. Evaluating for the learner, may mean that the effectiveness of the learning or the fit to curricular criteria may need to be assessed. The child as user will have typical learnability and ease of use/access goals similar to the adult.

APPENDIX 2: GUIDELINES FOR PEER TUTORING

Introduction

Peer tutoring is a usability evaluation method based on the idea that one child teaches another child how to use the software being evaluated in a social context familiar to them. The peer tutoring approach involves at least one tutor, a tutee, and an adult researcher called the *interactor* who guides the collaboration. The interactor does not operate the video camera or take notes, but teaches the tutors, guides the testing situation and asks questions during the peer collaboration efforts. Below you can find practical instructions and tips on how to apply the peer tutoring method in practice. The guidelines have been structured according to the phases of the evaluation process. Note that depending on your situation you can skip some of the steps presented.

Pre-test activities

1. Get context-sensitive. There are several variables that all affect the outcome of the evaluation that experimenters should be aware of:
1) *features of the interactive product* to be evaluated; who uses it, when and how it is used and what the stage of the product development is (single-user, multi-user, collaborative, task-oriented, entertaining, educating, home use, classroom use, mobile, fixed, PC, toy, multimodal, control devices, low-tech or high-tech prototype, etc.),
2) *children that are going to evaluate the product* (their abilities and developmental stage, age, gender, cultural and socio-economical background, energy level, expectations, fears, previous experiences, children's relationships, etc.),
3) *test environment* (school or home environment, indoors, outdoors, lighting, noise level, disruptions, the observational technology used, usability lab etc.),
4) *adults organizing the test* (their personality, behavior, language, roles, skills, etc.), and
5) *resources available* (time, people, schedules, money, school holidays etc.)

Tip! If you do not have much experience with working with children that belong to your target group, it is advisable to get to know this specific target group by interviewing parents, children or teachers, visiting a school or observing children in their freetime activities.

Tip! Play a lot with the product to be evaluated, and if possible, let some children use the product in an informal setting before making the final test plans.

Tip! It is good to visit the test location even before the pilot testing.

2. What do you want to know? What are your research questions and how can you measure them? It is not very fruitful to spend a lot of time organizing the test unless you know what kind of information you are interested in. So, you have to decide over and usually even prioritize your questions. You also need to decide what is the best method or combination of methods to find the answers to your questions. Here are a few examples of things you should also consider:

- Are you interested in longitudinal or one-time evaluations, i.e. how long will you work with the same children and how many children are needed to participate in the evaluations? How many times do you want to repeat the evaluation procedure?
- What are the features of the product you want to evaluate? This affects the task design and your presentation of the tasks to the children.

- What do you want to measure and what metrics are you going to use? Are you interested in usability flaws, task completion times, error frequency, fun, satisfaction, collaborative behavior, age and gender differences, language used etc.?
- Are you going to evaluate the product in a usability lab or in the field?
- What methods would you like to use? How much data will plain observation give and what survey techniques would augment the information obtained?

3. Decide what methods to use. The method or the combination of methods to choose depends on the issues mentioned above. It has been shown that some methods work better with specific age groups and with specific products (Van Kesteren et al., 2003). Peer tutoring is an option to consider in situations where:

- You are interested in how children learn and teach others how to use the product or when the software application to be tested is designed for social settings such as schools where it is beneficial that children can teach other children how to use the software
- You are interested in how children think and communicate about the system and when adult-child communication and power structures are sidetracking the test situation
- When you want your evaluations to be collaborative and they can be done in the children's own environment
- The cognitive load needs to be split – the tutees can concentrate on the task while the tutors handle the communication by teaching the tutee or by answering the adult's questions
- Children are active and communicate lively
- When you have time afterwards to analyze the video material obtained
- The design of a help system is based on children's views and experiences with the computer product
- You do not mind that children are given roles

4. Select the model for peer tutoring. Several variations of peer tutoring exist. Two different peer tutoring approaches can be used, depending on the number of tutors:
(1) In *two-on-one tutoring* (two children teach the third one), two children from a group of three are paired to participate in the tutor training phase. The third child (tutee) stays in the classroom with the rest of the class or in a separate room and is asked to join in later during the test session.
(2) In the *one-on-one tutoring* (one tutor teaches one tutee), a tutor or a group of tutors are first given an opportunity to use the product either alone or in a co-discovery manner and to learn to carry out the tasks. Then each tutor teaches his or her tutee how to use the product. The latter approach can also be applied as the *each-one-teach-one* mode, where each tutee acts as the tutor for the next child. The *one-on-one* approach is recommended if the children will be sitting in front of the computer.

Other modifications of the method are also possible, for example, cross-age tutoring. Peer tutoring can be combined with other methods as well.

5. Make a test plan. The test plan document will help you understand the whole evaluation process, share information and work tasks with your collegues and carry out the testing smoothly. In addition, it helps you re-organize similar testing later on.

6. Prepare for the unexpected. Usually interruptions and variations from the test plan inevitably ruin the test setup. However, with children you can't expect everything to go as planned. Think how you will react in unexpected situations and how you would guide the

situation towards the planned activity. Be also aware that children's external activities could tell you something about the product to be evaluated. In addition, you need to make sure that children can drop out of the evaluation session before it is completed without feeling they have done something wrong.

7. Plan your own behavior. It is very easy to get carried away if the children become very enthusiastic or bored with the test situation. Adults should not interfere with the children's activities but promote and guide their interactions and communication in the test situation.

Tip! Some people can naturally guide the test better than others - if you feel uncomfortable guiding the test, get someone else to do it.

Tip! Sometimes it is good to write down your lines, for example, how to introduce the test setup or how to give out tasks. Practice it in the pilot test.

8. Define the competence level between tutor and tutee, i.e. how and how long the tutor can use the interactive product before the tutoring session. Also, make sure that information is not passed from the tutor to the tutee before the tutoring.

9. Make the test arrangements and select the participants. Start making the test arrangement well in advance with the teacher and/or parents depending on where and when you are planning to conduct the testing. You should be flexible and respect the school's or day care center's timetables and holidays. Do not make it too difficult for the teacher to carry out her/his daily work and ask her/him to propose the time and the location for your tests. Also make sure that the test doesn't last longer than what the children can tolerate. Always meet the teacher before the actual testing if possible and visit the test location to make sure it fulfills your requirements. However, you should be aware that the teacher is not necessarily the best person to pick the participants for you. Sometimes they select participants that are the most difficult to deal with or do not get along very well. If the children are selected randomly, disruptive behavior might ruin the test setup. Make sure that you inform the parents and the children who are going to participate in the testing well in advance. And remember to plan the logistics and buy enough video cassettes or other recording media.

Tip! If two or more children are evaluating the product together, let them choose who they want to test the product with. Usually friends talk more with each other and feel more comfortable in the test situation.

Tip! One good option is to send a letter to the participants and their parents that contains all necessary information, a form to sign and a background questionnaire. This way you can ask the children and the parents for permission for the children's participation in the tests. Parents can also fill in the questionnaire with their children to provide you with background information on the children. In addition, the children can be asked to name two to three classmates they want to test the product with. The pairs can be formed according to the children's wishes.

10. Take care of ethical issues. Make sure that your research is ethically sound. This means that you take care of all necessary precautions such as the ones mentioned in the chapter "Ethical considerations when evaluating with children".

11. Always pilot test! The pilot test is your best friend. Even a small pilot test with one or two children will give you a lot of information on how to fine-tune and improve your test plan.

Setting up the test space

Reserve enough time for setting up the test space, especially when doing tests in the field. It is good to visit the test location well in advance to make plans on how to position the video cameras, chairs, computers or any other material. Sometimes you also need a room for the

participants to wait for their turn, so make sure that this space can be used for your purposes. It is advisable to do a final check with the test setup to see that everything works as planned. Usually it is also good to have a meeting with the teacher to confirm that timely matters are in order and that all the participants are attending school that day. The video camera(s) should be located and operated so that it does not distract the children. A good way is to put the video camera on record and then leave it without anyone looking after it. It is also recommended to place the children so that the tutor can not operate the computer and take over. A good way to make the test setup trouble-free is to give the tutor and tutee their own chairs to sit in.

During the test

The overall process of peer tutoring is presented below:

1. Introduce the roles and the test setup to the children.

First introduce yourself and tell children why you have come to their school or day care centre. Then describe the purpose of the evaluation to the children in general terms. You should also make children feel that they are helping you to make the product better and that they should not feel bad if the product doesn't work. Remember to tell the children that it is alright to quit at any time if they feel uncomfortable. It is important to give the children clear roles. Tutors need to know that it is their responsibility to teach and the tutees' to learn and use the product according to the tutor's instructions. Sometimes you can let the children to choose the roles by themselves. You can either explain the test setup to all participants at once or divide the children into smaller groups or pairs. Depending on the number of researchers participating in the study, the tutees could be in a separate room waiting for their turn, but this requires that an adult takes care of them. Another good solution would be to let the other children wait in their classroom and have their tutor pick them up. Note that depending on the type of user interface you are planning to evaluate the tasks and the introduction of the tutoring phase might need some modifications.

2. Training of the tutors. Depending on the peer tutoring approach you have chosen to use, you either let one tutor learn to use the product alone or have two to three tutors explore it collaboratively within the limits of the tasks given. It is important not to give out more information than necessary to prevent the children from imitating your instructions in their teaching. Usually co-discovery learning is a valuable tool in teaching the tutors. Once the tutors have been "trained", the actual tutoring session begins.

3. Tutoring session. In the tutoring session the tutor teaches his/her tutee how to carry out tasks. The trickiest part in peer tutoring is to design the tasks so that the tutor is able to remember what she/he is expected to teach. Note that very young children might only remember a part of the task at a time, so you might need to remind them to continue with their teaching. Children can be easily and gently reminded about their roles and tasks by using the *question asking protocol.* Two categories of questions can be used: (1) Questions that help tutors teach a tutee. When the tutor is teaching a tutee, the interactor does not give any instructions to the tutee, but asks a product or activity related question from the tutor when the tutor seems to need help or memory aid, or when the tutee is having serious trouble in carrying out the task. As mentioned earlier, the tasks can also be given in this manner, and (2) Comment related questions. When the tutor comments on an item in the user interface like "you made that funny sound!", the interactor asks a comment related question, for example, "how did that happen?", and the tutor replies "when you press that blue button".

Don’t be surprised if the tutor is more active than the tutee. It might also be surprising how much children actually try to show examples to others. Sometimes, however, tutors take over and start “operating the machine” rather than giving verbal instructions. In that case you need to have a “plan B” how to guide the tutor to only explain and, for example, point at objects on the screen, rather than carry out the tasks by themselves.

4. Wrapping up the session. Once the tutoring session is over and all tasks have been carried out, you can use a survey method or a combination of methods to obtain further information. Please refer to theory and guidelines related to survey methods in this tutorial and in the literature. During the wrapping up you and children can also discuss more informally and you can answer any remaining questions the children may have. You can also ask about the interesting behaviors you would like the children to explain. Remember to thank the children and all the other people for making your study possible!

After the test

Clear the test space and leave it as it was before you set the test up. Make sure that you keep the research material in a safe place and protect children’s privacy. Once the testing has been conducted successfully, it is time to analyse the recorded video material and all the other data collected. Since this is not a textbook on qualitative and quantitative research, the methods for analysing the data are left out. Usually teachers and some parents, and sometimes even children are interested about the findings of your study, so you can provide them a brief summary once you have finished working with the data. It is also very valuable to write down your experiences and suggestions for improvements on evaluation practices for later use. Remember to use your results!

Please cite the author (Johanna Höysniemi) when using this material.

Further information about peer tutoring can be found in:
Höysniemi, J., Hämäläinen, P., and Turkki, L. (2003). Using Peer Tutoring in Evaluating the Usability of a Physically Interactive Computer Game with Children. Interacting with Computers, Vol. 15/2, May 2003: Special Issue: on Interaction design and children. pp. 203-225.

Van Kesteren, I.E.H., Bekker, M.M.,Vermeeren, A.P.O.S., and Lloyd, P. (2003) Assessing Usability Evaluation Methods on their Effectiveness to Elicit Verbal Comments from Children Subjects. In Proceedings of Interaction Design and Children 2003, Preston, England, 1st – 3rd July, 2003. pp. 41-49.

APPENDIX 3: THE FUN TOOLKIT

This is an abridged version of results that have been published elsewhere. The full reference of the paper is: Read J. C., MacFarlane S. J., and Casey C. *Endurability, Engagement and Expectations: Measuring Children's Fun.* Presented at the Interaction Design and Children Workshop, Eindhoven (2002), Shaker Publishing Eds M.M. Bekker, P. Markopoulos, M. Kersten-Tsilkalkina pp 189 – 198 ISBN 90-423-0200-3. What is presented here is a set of tools that can be used with children to discover their opinions. Specifically, these have been used to measure 'how much fun' different computer interfaces were.

Tool 1 - The Funometer

This is essentially the 'Funometer' as developed by Risden, Hanna, & Kanerva, (1997). This tool has a vertical scale, which has been designed to simulate a thermometer, and which the children would use in a similar way, with a vertical bar representing the amount of fun. Figure 1 shows two funometers, one that is awaiting completion, and one that has been completed.

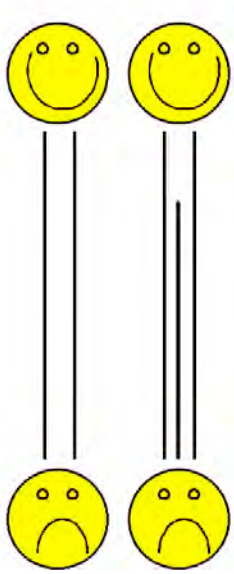

Figure 1. The Funometer – before and after completion

Tool 2 - The Smileyometer

This discrete variation of the Funometer was designed with the help of some children. It is based on a 1-5 Likert scale, and uses pictorial representations that the children contributed. Note that the neutral state does not have a face with a straight-line mouth. The weak smile was suggested by children, who reported that a straight line would depict anger. It is also interesting to note that the researcher had considered faces with two smiles or two frowns for the extremes, but the children just wanted bigger examples of the mouth articulations. Words were added to the faces, and the faces were presented in a horizontal row. Children were asked to tick one face.

Figure 2. The Smileyometer

It seems that the Funometer may be more useful with older children. If a number of different activities / events are being compared it is a mistake to present a row of Funometers or Smileyometers on a page, particularly when young children are using them. The Smileyometer was very easy to use, and the faces that have been presented are acceptable to children. Results from the 'Funometer' and the 'Smileyometer' have been shown to be very similar (N= 48, t = 0.99). This seems to indicate that one or the other but not both is needed. If either of these two tools is being used to compare activities, it is probably sensible to present one activity at a time in order to assist the child in discrimination.

Tool 3 - The Fun-Sorter

There are instances where the desire is to rank a series of connected activities, in order to establish which may be more appealing or which may be least fun. Repeated instances of the Funometer and the Smileyometer can be used but this variation on a repertory grid test (Fransella & Bannister, 1977) was devised. This Fun-Sorter has one or more constructs and a divided line (or table) that has as many spaces in it as there are activities to be compared. The children either write the activities in the spaces, or for younger children, picture cards can be made and placed on the empty grid.

	Best			Worst
Worked the best	B	D	A	C
Most Fun	D	A	B	C
Easiest to do	A			

Figure 3. A partially completed Fun-Sorter, for 4 activities and with 3 constructs

The Fun-Sorter seems to be most useful when comparing the relative merits of small numbers of activities (9 has been seen to be too many!). If only one construct is offered it can be used with young children particularly if picture cards are used to make the completion easy. Young children had difficulties discriminating between constructs like 'Easy to use' and 'Most fun' and children have been seen trying to make the grid 'fair' for the activities / events being compared. Given the difficulties that children may have with understanding constructs; it is important to describe these things using words and phrases that they know.

Tool 4 - Observation

Observing children can be very difficult and time consuming. Video footage can be used but this can be intrusive. Hidden cameras and multiple angles are difficult to set up in a school classroom, and bringing the child to a usability lab will change the nature of the trial. Children may move out of the line of vision of the camera or the hardware may fail.

The authors have measured engagement using video footage that was scored with reference to a set of positive and negative instantiations. The positive instantiations that were looked for were smiles, laughing, concentration signs (fingers in mouth, tongue out) excitable bouncing, and positive vocalization. Negative instantiations were frowns, signs of boredom (ear playing, fiddling) shrugs, and negative vocal instantiation.

Comparing the experiences of children who are working on one task with one another is problematic due to their differing animation. For one child, carrying out multiple activities, observational data may be quite valuable, in these circumstances, it is important to ensure that all the activities being presented offer the same opportunity for each of the positive and negative instantiations that are being recorded.

Tool 5 - Again-Again Table

This can be used to compare activities / events. The table lists some activities on the left hand side, and has three columns headed Yes, Maybe, and No. The child ticks either yes, maybe or no for each activity, having in each case considered the question 'Would you like to do this again?'

Would you like to do it Again?

	Yes	Maybe	No
Visit U Boat	✓		
Puppet show		✓	

Figure 4. Part of an Again – Again table

The Again-Again table gives results that are very similar to the Fun-Sorter. This suggests that it is fine to use just one of these two measures. Interestingly, children did not seem to feel the need to play fair on the Again-Again table, this may have been because it was about them and not about the activities; that is, the emphasis was different. This seems to indicate that one Again-Again table can be used to evaluate a number of activities. Where a large number of activities are being compared, the Again-Again table may be more useful than the Fun-Sorter. The advantage that the Fun-Sorter has is that it can be adapted to elicit other information, for instance how interesting or how easy an activity is, by adding relevant constructs.

Tool 6 – Remembrance

There is no real tool for this, simply a blank sheet of paper on which the child writes down what they did. It has been seen that children tend to remember the things they liked the most.

Summary

A sufficient measuring-fun toolkit would include the Smileyometer or the Funometer; an Again-Again table or a Fun-Sorter together with observations (where appropriate) and remembrance metrics.

Children often record events as ‘Brilliant’ and a strong motive for including observational data is that it may help discriminate. It is also important to note the findings relating to the desire on the part of the children to ensure ‘Fair Play’. These two traits seem to be common among children but it is likely that there comes a time when these traits become less pronounced as the children become ‘adult-like’ in their perceptions and views.

The difficulties that the younger children had with understanding and differentiating between constructs is also likely to change with maturity; although this is probably more ‘knowledge dependent ’ than the changes in the children’s perceptions which impinge on the Smileyometer scores.

The smileyometer / funometer can be used before an activity / event and it has been shown that children almost exclusively report having got what they expected to get! This suggests that fun measures may almost be carried out before an event took place! It is highly possible that the child’s perception of the fun they experience is governed by their expectations; in which case, the design of an interface is of much less importance than the ‘advertising and promotion’ that precedes the child’s experience of it.

For updates on this work visit www.chici.org
Please cite the author and the original reference when using this material.

APPENDIX 4: FEEDBACK

Content

Please answer the following questions in relation to the content that was delivered during the course:

Did we cover most of what you wanted in terms of content?

What would you like us to have covered that we did not cover?

What was covered that you felt could have been omitted?

Was the content covered in too little depth, in about the right depth or in too much?

Delivery

Please answer the following questions in relation to the way the course was run:

Was the oral presentation okay – do you have any suggestions for how it could be improved?

How easy was it for you to follow the AV materials? Any ideas for improvement?

How helpful do you expect the course notes will be?

Overall

Any other thoughts about this course or ideas on how to make it better?

:

REFERENCES AND BIBLIOGRAPHY

Some of these are cited in the notes, others are additional bibliography.

ACM/IEEE-CS Joint Task Force on Software Engineering Ethics and Professional Practices. *Software engineering code of ethics and professional practice.* Available at http://www.acm.org/serving/se/code.htm

Alborzi, H., Druin, A., Montemayor, J., Platner, M., Porteous, J., Sherman, L., Boltman, A., Taxen, G., Best, J., Hammer, J., Kruskal, A., Lal, A., Plaisant Schwann, T., Sumida, L., Wagner, R. and Hendler, J. (2000) Designing StoryRooms: interactive storytelling places for children *ACM SIGCHI 2000 Symposium on Designing Interactive Systems Brooklyn.* p95-104

American Psychological Association (APA). (2002) *Ethical principles of psychologists and code of conduct.* Available at http://www.apa.org/ethics/code2002.html

Barendregt, W., Bekker, M.M., Speekstra, M. (2003) *Empirical evaluation of usability and fun in computer games for children.* In Proceedings IFIP Interact '03, IOS Press, p705-708.

Bekker, M., Markopoulos, P., and Kersten-Tsikalkina, M. (2002) *Interaction Design and Children.* Shaker Publishing, Maastricht, The Netherlands.

Berkovitz, J. (1994) Graphical interfaces for young children in a software-based mathematics curriculum *CHI ' 94 Boston, Mass* ACM Press p247 - 248.

Blythe, M.A., Monk, A.F., Overbeeke, K. and Wright, P. (2003) *Funology: From Usability to Enjoyment.* Kluwer Academic Publishers, Amsterdam.

Boren, T. and Ramey, J. (2000) Thinking aloud: reconciling theory and practice, *IEEE Transactions on professional communication,* 43, (3).

Boucher, J. and Osgood, C. E. (1969) The Pollyanna hypothesis, *Journal of Verbal Learning and Verbal Behaviour,* 8, 1-8.

Breakwell, G. (2000) Interviewing, in *Research Methods in Psychology* (Eds, Breakwell, G., Hammond, S. and Fife-Shaw, C.) Sage Publications, London, p245 - 246.

Bruckman, A., and Bandlow, A. (2003) HCI For Kids. in Jacko, J., and Sears, A., (Eds.), *Human-Computer Interaction Handbook,* Lawrence Erlbaum Associates, p428–440.

Buckleitner, W. (1999) The state of Children's Software Evaluation-Yesterday, Today And In the 21st Century, in *Information Technology in Childhood Education* (Ed, Warren, B.), p211 - 220.

Burmeister, O. K. *HCI Professionalism: Ethical concerns in Usability Engineering.* In proceedings of Research and Practice in Information Technology (2001) Vol 1, 11-17.

Cassell, J., & Ryokai, K. (2001). Making space for voice: Technologies to support children's fantasy and storytelling, *Personal Technologies,* 5 (3), 203-224.

Childnet International (2002) *Safety Advice – Web Design*, http://www.childnet-int.org/safety/webdesign.html

Corsaro, W.A, (1992). Interpretive reproduction in children's peer cultures, *Social Psychology Quarterly,* 55, 160-177.

Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches. In N. M. Webb (Ed.), *Peer interaction, problem-solving, and cognition: multidisciplinary perspectives*. New York: Pergamon Press, 9-19.

Dix, A., Finlay, J., Abowd, G., Beale, R., (2004) *Human – Computer Interaction* Prentice Hall,

Donker, A., and Markopoulos, P., *A comparison of think-aloud, questionnaires and interviews for testing usability with children*, In: Faulkner, J.Finlay, F.Detienne, (Eds.) "People and Computers XVI - Proceedings HCI 2002", Springer, p305-316.

Draper, S. W. (1999) Analysing fun as a candidate software requirement, *Personal technology*, 3, (1) p1-6.

Druin, A. (Ed.) (1999) *The Design of Children's Technology*, Morgan Kaufmann Publishers, Inc.

Druin, A. & Solomon, C. (1996) *Designing Multimedia Environments for Children*, John Wiley and Sons Ltd.

Druin, A., Bederson, B., Boltman, A., Miura, A., Knotts-Callaghan, D. & Platt, M. (1999) Children as our technology design partners, in *The Design of Children's technology* (Ed, Druin, A.) Morgan Kaufmann, San Francisco, CA, p51 - 72.

Dumas, J.R., & Redish, J.C., (1993) A practical guide to usability testing. Ablex Publishing Corporation, Norwood NJ.

Erikson, E. (1950) *Childhood and Society*, Norton, New York

Erikson, K.A. and Simon, H.A. (1984) Protocol Analysis: Verbal Reports as Data. MIT Press.

ETSI, *Access to ICT by young people: Issues and guidelines.* ETSI DTR/HF-0034 v.0.0.5. http://www.etsi.org.

Fine, G. A. & Sandstrom, K. L. (1988) *Knowing Children: Participant Observation with Mirrors*, Sage Publications, Bevely Hills.

Garvey, C. (1986). In E.C. Mueller & C.R. Cooper (Eds.), *Process and outcome in peer relationships*, Academic Press, Inc., Orlando, p329-345.

Gaustad, J. (1993). Peer and cross-age tutoring, *ERIC Digest*, 79.

Gell, A. (1992) The technology of Enchantment and the enchantment of technology, in *Anthropology, Art and Aesthetics* (Eds, Coote, J. and Shelton, A.) Clarendon Press, Oxford.

Gilutz, S. & Nielsen, J., (2002) Usability Websites for Children: 70 Design Guidelines. Nielsen Norman Group. http://www.NNgroup.com/reports/kids.

Goodblad, S., & Hirst, B. (1989). *Peer tutoring, a guide to learning and teaching*, Kogan Page, London.

Hämäläinen, P., & Höysniemi, J. (2002). *A Computer Vision and Hearing Based User Interface for a Computer Game for Children.* In Proceedings of the 7th ERCIM Workshop "User Interfaces For All", 23-25 October 2002, Paris.

Hanna, L., Risden, K., Czerwinski, M. and Alexander, K., J (1999) The Role of Usability Research in Designing Children's Computer Products, in *The Design of Children's Technology* (Ed, Druin, A.) Morgan Kaufmann, San Francisco, p4-26.

Hanna, L., Risden, K. & Alexander, K., J (1997) Guidelines for usability testing with children, *Interactions*, (5) 9-14.

Haugland, S. (2000) Computers and Young Children.

Höysniemi, J. and Hämäläinen, P. (2003). Who is afraid of spiders? Two Perceptive Computer Games for Children. *In Proceedings of International Conference on Interaction Design and Children*, 1-3 July 2003, Preston, UK.

Höysniemi, J., Hämäläinen, P., and Turkki, L. (2002). Extended abstract: Using Peer Tutoring in Evaluating the Usability of a Physically Interactive Computer Game with Children. *In Proceedings of International Workshop on Interaction Design and Children*, 28-29 August 2002, Eindhoven, The Netherlands.

Höysniemi, J., Hämäläinen, P., and Turkki, L. (2003). Using Peer Tutoring in Evaluating the Usability of a Physically Interactive Computer Game with Children. *Interacting with Computers*, Vol. 15/2, May 2003: Special Issue: on Interaction design and children. pp. 203-225.

Inkpen, K. (2001) Drag-and-Drop versus Point-and-Click Mouse Interaction Styles for Children, *ACM Transactions on Computer-Human Interaction*, 8, (1) 1-33.

Jones, M. G. (1998), Creating Engagement in Computer-based Learning Environments, Vol. 2000 IT Forum.

Kafai, Y. (1999) Children as Designers, Testers, and Evaluators of Educational Software, in *The Design of Children's Technology* (Ed, Druin, A.) Morgan Kaufmann, San Francisco, pp. 124 - 145.

Kafai, Y., Ching, C. C. and Marshall, S. (1997) Children as Designers of Educational Multimedia Software, *Computers Education*, 29, (2/3) 117-126.

Kemp, J. A. M. and van Gelderen, T. (1996) Co-discovery exploration: an informal method for iterative design of consumer products, in *Usability Evaluation in Industry* (Eds, Jordan, P. W., Thomas, B., Weerdmeester, B. A. and McClelland, I. L.) Taylor and Francis, London, p139 - 146.

Kersten-Tsilkalkina, M. and Bekker, M. (2001) Evaluating usability and pleasure of children's products *International Conference on Affective Human Factors Design Singapore*.

MacFarlane, S., Nicol, T., Read J. and Snape, L. (Eds.) *Proceedings of Interaction Design and Children 2003*, Child Computer Interaction Group, University of Central Lancashire, Preston, UK.

Malone, T. W. and Lepper, M. R. (1987) Making learning fun: A taxonomy of intrinsic motivations for learning, in *Aptitude, Learning and instruction* (Eds, Snow, R. E. and Farr, M. J.) Erlbaum, Hillsdale, NJ.

Markopoulos, P. and Bekker, M. (2003) On the assessment of usability testing methods for children, *Interacting with Computers*, 15(2), p227-243.

Molich, R., Laurel, B., Snyder, C., Quesenberry, W., & Wilson, C.E., *Panel: Ethics in HCI.* CHI '01 extended abstracts on Human factors in computer systems 2001, Seattle, Washington.

Nielsen, J. (1993). *Usability engineering*, Academic Press, Boston.

Nielsen, J. (1994) Heuristic Evaluation, in *Usability Inspection Methods* (Eds, Nielsen, J. and Mack, R. L.) John Wiley.

Oliver, R. L. & Burke, R. R. (1999) Expectation Processes in Satisfaction Formation, *Journal of Service Research,* 1, (3) 196 - 214.

Passig, D. and Levin, H. (1999) Gender Differences with multimedia learning interfaces, *Computers in Human Behavior*, 15, (2) 173 - 183.

Piaget, J. (1929). *Introduction to the child's conception of the world.* Harcourt, New York

Price, S., Rogers, Y., Scaife, M., Stanton, D. & Neale, H. (2003) Using 'tangibles' to promote novel forms of playful learning. *Interacting with Computers, Special Issue: on Interaction design and children* Vol. 15/3, May 2003, pp 169-185.

Read, J. C. (2002) Optimising Text Input for Young Children using Computers for Creative Writing Tasks *HCI 2002* London, England

Read, J. C. & MacFarlane, S. J. (2000) Measuring Fun *Computers and Fun 3* York, England

Read, J. C., Gregory, P., MacFarlane, S. J., McManus, B., Gray, P. and Patel, R. (2002) An Investigation of Participatory Design with Children - Informant, Balanced and Facilitated Design *Interaction Design and Children* Eindhoven

Read, J. C., MacFarlane, S. J. and Casey, C. (2000) 'Postcards to Grandad'- A study of Handwriting Recognition with Young Children *HCI 2000 Sunderland,* England 2 51-52

Read, J. C., MacFarlane, S. J. and Casey, C. (2001) Can Natural Language Recognition Technologies be used to enhance the Learning experience of Young Children? *Computers and Learning* Warwick, UK

Read, J. C., MacFarlane, S. J. and Casey, C. (2001) Expectations and Endurability - Measuring Fun *Computers and Fun 4* York, England

Read, J. C., MacFarlane, S. J. and Casey, C. (2001) Measuring the Usability of Text Input Methods for Children *HCI2001* Lille, France Springer Verlag 1 559 - 572

Read, J. C., MacFarlane, S. J. and Casey, C. (2002) Designing a Handwriting Recognition Based Writing Environment for Children *8th International EARLI SIG Writing Conference* Staffordshire, England

Read, J. C., MacFarlane, S. J. and Casey, C. (2002) Endurability, Engagement and Expectations: Measuring Children's Fun *Interaction Design and Children* Eindhoven

Read, J. C., MacFarlane, S. J. and Casey, C. (2002) Oops! Silly me! Errors in a Handwriting Recognition-based Text entry Interface for Children *NordiChi 2002* Aarhus, Denmark

Read, J. C., MacFarlane, S. J. and Casey, C. (2002) Pens Behaving badly - Usability of Pens and Graphics Tablets for Text Entry with Children *ACM UIST'02* Paris, France

Read, J. C., MacFarlane, S. J. and Casey, C. (2003b) A Comparison of two on-line Handwriting Recognition Methods for Unconstrained Text Entry by Children *HCI 2003 Bath* Research Press International, Bristol 2 29 - 32

Read, J. C., MacFarlane, S. J. and Casey, C. (2003c) 'Good enough for what? Acceptance of Handwriting Recognition Errors by Child Users *IDC 2003 Preston, UK* ACM Press 155

Read, J. C., MacFarlane, S. J. and Casey, C. (2003d) What's going on? Discovering what Children understand about Handwriting Recognition Interfaces *IDC 2003 Preston, UK* ACM Press 135 - 140

Risden, K., Hanna, E. and Kanerva, A. (1997) Dimensions of intrinsic motivation in children's favorite computer activities *Society for Research in Child Development* Washington, DC

Rode, J. A., Stringer, M., Toye, E.F., Simpson, A.R. and Blackwell, A.F, (2003) Curriculum Focused Design. In MacFarlane, S., Nicol, T., Read J. and Snape, L. (Eds.) Interaction Design and Children 2003, Child Computer Interaction Group, University of Central Lancashire, Preston, England, 119-126.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context.* Oxford University Press, Oxford.

Scaife, M., Rogers, Y., Aldrich, F. and Davies, M. (1997) Designing For or Designing With? Informant Design for Interactive Learning Environments. *CHI 97* Atlanta p343-350

Schuler, D. and Namioka, A. (Eds.) (1993) *Participatory Design: Principles and Practices,* Lawrence Erlbaum, Hillsdale, NJ.

Sedighian, K. and Sedighian, A. S. (1997) Aesthetic Response: Children's Reactions to Color and Graphics in Educational Software *ED-MEDIA97: World Conference on Educational Multimedia and Hypermedia* Calgary

Taxen, G., Druin, A., Fast, C. and Kjellin, M. (2001) KidStory: a technology design partnership with children, *Behaviour and Information Technology,* 20, (2) 119 - 125.

Topping, K. (1988). *The peer tutoring handbook: promoting co-operative learning.* Billing & Sons Limited, Worcester, Great Britain.

Uden, L. and Dix, A. (2000) Iconic Interfaces For Kids On The Internet *IFIP World Computer Congress Beijing* p279 - 286

Usability SIG (2002) *Design and Evaluation with the Old and the Young,* http://www.stcsig.org/usability/topics/special.html

Vailliancourt, P. M. (1973) Stability of children's survey responses *Public Opinion Quarterly,* 37, 373 - 387

Van Kesteren, I.E.H., Bekker, M.M.,Vermeeren, A.P.O.S., and Lloyd, P. (2003) Assessing Usability Evaluation Methods on their Effectiveness to Elicit Verbal Comments from Children Subjects. In MacFarlane, S., Nicol, T., Read J. and Snape, L. (Eds.) *Proceedings of Interaction Design and Children 2003,* Child Computer Interaction Group, University of Central Lancashire, Preston, UK. 41-49.

Wartella, E., O'Keefe, B., and Scantlin, R., (2000) Growing up with interactive media: what we know and what we don't know about the impact of new media on children. A Report to the Markle Foundation, available on-line at http://www.hcibib.org/kids/.

Wilson, C. & Fleek, F., (2001) *Ethics in HCI and Usability,* http://www.stcsig.org/usability/topics/ethics.html

THE SLIDES