On the assessment of usability testing methods for children

Panos Markopoulos*, Mathilde Bekker

Technology Management, Technical University of Eindhoven, Den Dolech 2, 5600 MB Eindhoven, The Netherlands

Abstract

The paper motivates the need to acquire methodological knowledge for involving children as test users in usability testing. It introduces a methodological framework for delineating comparative assessments of usability testing methods for children participants. This framework consists in three dimensions: (1) assessment criteria for usability testing methods, (2) characteristics describing usability testing methods and, finally, (3) characteristics of children that may impact upon the process and the result of usability testing. Two comparative studies are discussed in the context of this framework along with implications for future research.

© 2003 Elsevier Science B.V. All rights reserved.

1. Introduction

The field of human computer interaction has in the last few years shown an increasing interest in generating scientific and methodological knowledge about how to design interactive systems for children. This paper focuses on usability testing. It reports the set-up and the initial results of a research effort to generate sound methodological knowledge for involving children as participants in usability testing. Currently, no report exists of a systematic comparison of usability testing methods (UTMs) focusing on children users. Neither does a systematic effort exist to specify the method and instrumentation of usability testing when test participants are children.

A rich and lively debate has developed in the field of Human Computer Interaction around the topic of comparing usability evaluation methods for adult users. We draw lessons from this more general debate and contribute to it with a framework for relating such studies that helps target issues arising when usability test participants are children.

* Corresponding author.
E-mail addresses: p.markopoulos@tue.nl (P. Markopoulos); m.m.bekker@tue.nl (M. Bekker).
Section 2 summarises the state of the art concerning the involvement of children in the evaluation of interactive technologies. Section 3 examines briefly comparative studies of usability evaluation methods, providing some entry points to the relevant literature. Section 4 outlines a framework for characterising assessments of usability testing methods for children. Sections 5–7 describe the dimensions of the framework in more detail: criteria for assessing UTMs, characteristics for describing usability testing methods and characteristics of children that may impact usability tests. Section 8 illustrates how research questions can be derived from the framework with two case studies. Section 9 looks critically at the problems that the framework leaves open and the problems that remain to be addressed by future research.

2. On the involvement of children in usability testing

Gradually, children are recognised as useful and active participants in the design of interactive systems for children. Druin (2002) has examined the roles that children can assume in the design of technology advocating their active participation as design partners. Such an active participation by children promises to tap into their creativity and maximise the benefits of their involvement, but it may not always be practical or accepted within the context of a design project. On the other hand, a baseline requirement for user centred design of children’s technology is that children act as testers (test-users) of technology intended for children. This is certainly not novel in itself. There is a long tradition of involving children as testers of products but there has been (to date) no systematic methodological investigation of how such usability testing should be conducted.

Bruckleitner (1999) reviews the current state-of-the art for children’s software evaluation. The evaluation practices Bruckleitner describes are predominantly expert reviews intended for recommending software to educators and to parents. Such evaluations are based on expert review rather than testing with children. Moreover, as Bruckleitner comments, very few studies have attempted to compare the assessments of experts to those of children.

Usability testing practices involving children that have been reported in the research literature are very diverse. These practices vary from simply asking children what was successful and what not, or asking what were the ‘bugs’ of the software tested, to analysing the stories children make with a story telling program. No systematic account of which testing practices work and which do not has been made. Druin (2002) reports that children are ‘incredibly honest and at times harsh in their assessment of technology’. She argues that children can readily slip into the role of the tester, because they are asked simply to use technology, which is something they do anyway. However, when children are involved in usability testing they are required to do much more than use the technology under evaluation: they have to adapt to a testing environment, interact with the facilitator, follow some processes and, usually, contribute to the evaluation by reporting on their experiences. Consequently, there are several issues that need to be considered for conducting a usability test with children. Druin (2002) suggests that short testing sessions, e.g. one afternoon, are found to give useful input. This is useful advice but rather coarse-grained. The usability tester who has less experience with children, still has to figure out
which method to use, what results to expect and to wonder how to adapt (if at all) a UTM to the special characteristics of children.

Hanna et al. (1997) provide the most comprehensive treatment of the subject to date. They describe a set of guidelines for usability testing with children in the laboratory. This set of guidelines touches on a wide diversity of subjects, such as the decoration of the room, the observation equipment used (e.g. use small microphones), the behaviour of the experimenter (how to encourage them when they fail at a task), and the varying capabilities of children of different age groups. For example, they suggest giving short tasks to younger children, that children above 14 act as adults in a testing session, etc.

The attention to detail paid by Hanna, Risden and Alexander and the wide range of contextual factors that acquire importance when children are brought into the usability laboratory reveal the complexity of the topic and the fragile nature of the process. Their advice seems rational and very useful to practitioners. However, it is not very clear on what evidence it is based and thus we conclude it describes the personal experience of the researchers involved. The research described here aims to add to such experience-based advice with a systematic account of how the different capacities and characteristics of children affect usability testing. The objective is to derive empirically methodological knowledge for prescribing how to conduct usability testing with children participants. Several questions need to be addressed to meet this objective: What makes children special as testers of technology? What capabilities/limitations of them need to be taken into account? How can we accommodate their different skills and interests? To our knowledge there has been no research reported to the time of writing this article, concerning the assessment of usability evaluation methods when the test subjects are children.

3. Comparing usability testing methods

Jorgensen (1999) attempts to put some order in the prolific amount of research concerning usability evaluation methodology. He characterises research contributions in four levels:

1. Developing, validating and documenting an individual evaluation method.
2. Developing a conceptual and methodological foundation for usability evaluation methods.
3. Comparative studies, e.g. those reviewed by Gray and Salzman (1998).
4. Consolidation of comparative studies to a single body of cumulative evidence concerning the effectiveness and efficiency of evaluation methods.

There is almost no work addressing the latter level. The research effort outlined here attempts to fill this gap for the particular user group targeted. It aims to compare evaluation methods (more specifically usability testing methods) not as individual instances, but by referring to their conceptual and methodological foundations, i.e. the elements populating level 2. This approach paves the way for cumulative development of knowledge at the fourth level of Jorgensen.
During the 90s numerous publications reported comparisons of usability testing with usability inspection methods. A shared objective of such comparisons was to gauge the validity of the inspection methods and to establish cheaper methods for usability evaluation than user testing. The practical utility of these studies was to help ensure results that are valid, reliable and come at a low cost. For example, Nielsen and Landauer (1993) provided a mathematical model for predicting the number of usability problems identified by a single evaluator as a function of the number of users analysed. Jacobsen (1999) proposed a similar formulation for the number of evaluators who analyse the collected data. Clearly, such studies concern many of the parameters that define a usability evaluation and generalisation of their results should be done with caution. Our aim is to examine whether and how such results extend to children test-users. We are interested in isolating and understanding the effect of variations in the usability test set-up for the cases where the test users are children.

Comparative studies of UTMs here have had a significant impact on shaping usability evaluation practice. Gray and Salzman in their seminal paper (Gray and Salzman, 1998) offer a critical review of a set of comparative studies. Their overriding conclusion is that there are serious threats to the validity of the studies they reviewed which have not been sufficiently addressed and many of the conclusions drawn from the comparisons are undue generalisations. Reviewers of the paper who answered back on the same journal issue did not in essence disagree with the criticisms but, rather, asked for such criticisms to be kept in perspective. The practical applicability of the results necessitates flexibility in the experimental design and the interpretation of its results. Recent research by Sears (1997) and Hartson et al. (2001) are starting to introduce structure and rigour into this sub-field addressing some of the methodological concerns of Gray and Salzman.

4. A framework for comparing usability testing methods for children

This section describes a framework for relating comparative studies of UTMs with respect to their appropriateness for children. The purpose of the framework is not to guide a classification or even a meta-analysis of this type of studies as, so far, there have not been previous studies in this domain. Rather, the framework serves to structure this problem space in a manner that allows links to be drawn across studies and their results to accumulate. The proposed framework has three dimensions:

- The first dimension accommodates criteria for assessing a UTM. A common criterion for many comparative studies is the sum of usability problems reported by a group of subjects. This is a simple and practical measure that partially gauges the effectiveness of the UTM. A closer examination raises several questions: Are all usability problems reported actually usability problems? Are we sure we don’t count the same problem twice? Are all problems equally serious? How can we optimise the resources needed for a usability test?
- The second dimension consists of a few characteristics of UTMs which together help define any particular method operationally, i.e. in sufficient detail to guide a practitioner to apply them. These attributes could be, for example, how many users are involved in
one session, what do they do, what does the facilitator do, etc. The intention is that
comparative studies should assess the impact upon the performance of a UTM
(assessed with the criteria of the first dimension) that may be attributed to choices or
variations of the UTM in terms of these attributes. Rather than comparing UTMs in the
form with which they have been originally introduced in the research literature or in the
form found most appropriate by practising professionals, comparative studies that
position UTM along this dimension can claim higher construct validity.

Considering the special target user group that this research focuses on, an obvious
question emerges. How does the participation of children influence the UTM and how
should comparative studies be oriented towards generating methodological guidelines for
involving children in usability testing.

- The third dimension concerns characteristics of children, i.e. evolving interests and
capabilities of children that affect their performance as test participants and the
results of the usability test where they participate. Such a characteristic is, for
example, the ability of children to verbalise in order to express likes or dislikes,
reporting what they find problematic in using a system, etc. Looking at children in
terms of these characteristics helps generate more refined and accurate advice than
treating all children uniformly.

See Fig. 1 for an overview of the dimensions and their associated ‘attributes’. The
dimensions of the framework are examined in more detail in the following sections.

5. Criteria for assessing usability testing methods

The first dimension of the framework consists of several criteria against which usability
testing methods can be assessed and that are traditionally used to describe methods in
general. For example, practitioners need to know:

- Will they be able to apply a particular UTM for their problem (robustness)?
- How good results does this UTM produce (effectiveness)?
- How expensive it is to apply a UTM in terms of time or other resources (efficiency)?

Hartson et al. (2001) propose quantified criteria for assessing usability evaluation
methods (i.e. both inspection and testing techniques). To enable links to that work much,
but not all, of their terminology is adopted. Notably, effectiveness is considered here not as
a singular numeric expression but, rather, as a vector that describes UTM results through
more detailed criteria such as validity, reliability and thoroughness. These criteria are
explained below. The definitions below refer to UTMs but they extend directly to usability
evaluation methods in general.

Robustness. The robustness of a method concerns the feasibility of applying it to
different contexts, such as the laboratory versus the field, applying it to different products,
such work-related versus entertainment-related, early in product development versus late, etc. The robustness of a method can be assessed through trials in different domains. In the special case of children testers, robustness pertains e.g. to whether different ages, skills or capacities of children should render inapplicable a method established for adult participants.

For example, children under seven may have difficult time reading instructions or filling-in written questionnaires. A systematic treatment of this topic calls for explicating the requirements UTMs set upon test participants and encoding the intended scope of application of each method, e.g. type of product, part of the development cycle, age appropriateness, etc. Case studies such as that summarised in Section 8.1 are needed to establish the robustness of usability testing methods with children users.

**Reliability.** The reliability of a usability testing method pertains to whether factors external to the method influence its outcome when the same testing procedure is used for the same product. Typical factors that might influence the reliability of UTMs are
the differences between individuals acting as test participants or as facilitators (cf. Jacobsen, 1999). The outcomes compared in such studies are typically the set of problems identified. Being able to identify to a large extent the same set of problems regardless of the actual users or facilitators testing a product is a baseline requirement for any usability testing method. Otherwise the method offers no guarantees for improving usability of the eventual product in real use. Reliability is improved with higher thoroughness (a criterion discussed below in relation to effectiveness).

Validity. Validity is an important descriptor of effectiveness. The validity of a usability testing method pertains to whether the problems it helps uncover are actually usability problems or not. Such a view would suggest that there is an unequivocal definition of what constitutes a usability problem or at least way of deciding whether an issue named as a usability problem really is one or not (an equivalent to a ‘Litmus’ test in Chemistry). Such definitions that would help judge the ‘realness’ of an identified problem do not exist and comparisons between methods can only compare the overlap of different procedures for identifying usability problems without knowing which one is closer to the ‘truth’.

Hartson et al. (2001) discuss various solutions to this problem and their respective limitations. They lean towards relying on the judgement of end-users as to whether usability problems identified through evaluation are real or not. More research and more convergence among researchers in this field is needed to establish benchmarks appropriate for comparing usability tests. Such benchmarks should identify representative products to which usability tests should be applied and a process for creating and maintaining ‘standard sets’ of known problems for such benchmark products.

Thoroughness and problem counting. Thoroughness aims to describe the proportion of all usability problems of a product that are found through a test. In practice this assumes the ability to enumerate all usability problems of a product, so it can only be approximated, e.g. by the union of problems identified through a combination of methods (which implicitly assumes that all problems are real and adopt the union of problems found as the ‘complete’ set).

Counting usability problems also needs to be approached with care. Different usability tests and comparative studies should share a common definition of what a usability problem is. Duplication of problems needs to be eliminated and this requires the definition of classification schemes for usability reports (cf. Cockton and Lavery, 1999). At the very least researchers should precise what they count and draw attention to the pitfalls surrounding comparisons.

Efficiency. Efficiency measures can be the amount of the resources used in relation to the outcome, e.g. the number of real problems found per child participant. The studies of (Nielsen and Landauer, 1993) and Jacobsen (1999) present mathematical models of the efficiency of a method with respect to the number of test participants and in the latter case of usability evaluators.

The above list of criteria for UTMs is not yet comprehensive, but provides sufficient means for the purposes of relating comparative assessments of UTMs to each other. The discussion above helps justify some of the choices made for the case studies in Section 8. It also points to the need for standardising definitions of what is a usability problem,
procedures for classifying such problem reports and, eventually, for benchmarking usability testing results.

### 6. Characteristics of usability testing methods

The second dimension of the framework covers various characteristics for describing UTMs. This dimension provides a common schema for describing UTMs. To safeguard construct validity, comparative studies of UTMs need to both refer to the various instantiations of a particular method in the literature and make explicit how these variables are set.

**Participants**. UTMs are characterised by the number of participants involved, whether they come alone, in pairs or larger groups, how representative they are of the target user group, their motivation etc. Many methods involve a single user with a single experimenter.

**Facilitator and evaluator**. The facilitator runs the usability test while the evaluator is the one who analyses the data. These roles can be assumed by the same person.

**Environment/context**. Is it in the laboratory, in the real working environment, in the classroom? Is this place noisy, are there more people around, or is it isolated?

**The interaction tasks**. During a usability-test participants are given a set of tasks to perform. These may be abstract, e.g. to type-in a short sentence, or, concrete, e.g. to type-in ‘ABCDEF’. Tasks can be described by their complexity, their abstractness, their number and duration, whether they are exploratory or task-oriented (Vermeeren, 1999).

**Capture of data**. What technique is used to capture empirical data? Are they explicit tasks by the user or is the user observed? What technique is used: e.g. interview, logging, questionnaire, video recording or note keeping? Do participant tasks aiming for data collection interfere with primary interaction tasks?

**Procedure**. Do subjects work alone, in pairs, or in a group? Does the facilitator interact with them, observe them, interview them, and prompt them? What instructions are given? When? In what form?

Arguably the type of artefact tested impacts the usability test significantly. However, the artefact is external to the operational definition of the UTM so the performance of a UTM for different artefacts is, for our purposes, a robustness issue. In this case, robustness concerns the degree to which the outcome of the UTM is affected by whether it is a learning application or a non-educational game. The purpose of a usability test is also considered here as orthogonal to its operational definition, but may have significant impact on the utility of a UTM in a particular context. E.g. some UTMs may be less effective than others for generating information about the detailed ‘look and feel’ rather than the role of the artefact in the daily life of its users.

### 7. Characteristics of children that impact the usability test

The third dimension describes characteristics of children that might influence the usability test. This section does not attempt to provide a comprehensive profile for children
as technology users nor does it attempt an overview of relevant theories in child development. Rather, the aim of the exposition is to select a few characteristics of children that are expected to impact the conduct and the outcome of a usability test. Relevant hypotheses may then be derived and tested empirically. These are the following:

*Capacity and inclination to verbalise.* Children have developing capacity to verbalise (both vocally and in writing). Data collection methods used to identify usability problems might be influenced by the fact that children have different verbalisation capabilities. A related skill is the ability to think aloud. This asks for the child to translate their experiences to verbal statements. According to Hanna et al. (1999) children below the age of 12 are likely to be unable to think aloud. Our studies (see Section 8) seem to go against this rule of thumb. Some children may not be used to speaking up to adults and may be less likely to report usability problems. Extroversion and verbalisation skills are thus important independent variables to control. It is reasonable to expect that these capabilities of children have direct impact on the outcome of the usability test, so they should be assessed directly, rather than indirectly through the age of the children.

*Capability to concentrate.* Children have developing capability to concentrate to a single activity and to pursue tasks. Consequently, tasks of different complexity and size should be given to different ages. Even though Hanna et al. (1997) mention that children can concentrate for about 30 min, our experience is that children (9–12) who are enjoying themselves have no problem with participating in longer sessions (in some cases, after 45 min, they were eager to continue).

*Children’s motivation.* Another factor that may influence the outcome of usability testing is children’s motivation. For example, frequent intervention by the facilitator together with high motivation to please adults (evident at younger ages), could also influence the outcome of usability testing procedures (Hanna et al., 1997).

*Ability to adjust to strange environments and surroundings.* Children have varying abilities to adjust to strange environments and surroundings. This can be important when testing is done at a strange location such as a usability laboratory. The social environment for the usability test might also seriously distract children, e.g. Hanna et al. (1997) mention that their peers, noise, other objects in the environment, might distract them while having their parents around is advisable. Allowing them some time to become used to the environment will minimise the amount of distraction.

*Trustworthiness of self-report.* Children are reported to be very honest (Druin, 2002) 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 looks for them or they may conceal problems if they think they would offend the software creators. Depending on their age children are influenced by other children, parents and teachers, etc., in different ways (Acuff and Reiher, 1997). Over time they shift from needing approval by their peers to developing an increased independence from adults. These processes can influence what bias may occur because of the age and roles taken by facilitators and others present at a usability test. The biasing effect by the facilitator, which has been studied previously for adult testers, cf. Bentley (2000), can potentially be larger or different for children users.
Ability for abstract and logical thinking. Children develop the ability for abstract and logical thinking over time (Berk, 1997; Acuff and Reiher, 1997). They also become better at doing more complex reasoning, such as cause and effect reasoning. This influences their ability to understand abstract task descriptions and abstract questions for feedback. It also influences their ability to compare or judge products on multiple aspects, varying in level of detail. Furthermore, it influences the number of concepts they can keep in mind at once (Greig and Taylor, 1999). Other method dimensions, such as concurrent verbalisation and marking along Likert scales have to be assessed for appropriateness. Hanna et al. (1999) propose an adaptation of Likert scales, but there is no evidence that this will work sufficiently or not. Clearly, producing methodological advice for usability testing should investigate the effect of age on the ability of children to carry out some of these tasks.

Monitor progress towards a goal. Children slowly develop the ability to monitor progress towards a goal over childhood and adolescence (Berk, 1997). They have to build up the ability to check outcomes of a task, and learn to redirect unsuccessful efforts. The level of this ability will influence how good they are at conducting goal directed tasks and assessing whether they have done them correctly.

Gender differences. These develop and change as children become older (Berk, 1997). In two of our experiences involving children between 9 and 11 years old, girls have been more verbose than boys and girls provided more arguments for their opinion. Furthermore, girls and boys are likely to have different evaluation criteria for assessing products, because of their preferences for different kinds of products and activities (Acuff and Reiher, 1997).

Knowledge of language and concepts. Since children’s knowledge of language and concepts is developing, appropriate language or vocabulary is important to ensure that children understand what is expected from them. Using age and culture appropriate language will influence children’s understanding of the usability test procedure in general, and more specifically the task descriptions, prompting instructions and requests for feedback.

Knowledge and skills. These are directly related to the usability of a system or to the tasks given to the test-participant for the purposes of collecting data. For example, consider prior experience with computing. Wartella et al. (2000) say that children owning a computer demonstrate more positive attitudes, more enthusiasm and report more self-confidence and ease when using computers than those who do not. This may affect the outcome of a usability test and confound analyses of the age appropriateness of various usability testing processes.

8. Deriving research questions from the framework

Comparative studies of UTMs can be described along the first two dimensions and potential complications resulting from children-participants are covered by the third. Fig. 1 shows how research hypotheses are placed within this framework. Variations on characteristics of children or variations on a UTM attribute can be hypothesized to relate to the robustness, effectiveness or efficiency of a UTM. Clearly, such hypotheses are interesting to test if they can potentially lead to guidelines for setting up a usability test.
For example, varying the degree of intervention by the facilitator, e.g. in frequency, length and tone (positive/negative) may be hypothesized to affect children testers more than adults. This effect might depend on gender and/or age of the children. The validity of the usability problems found is then the most relevant attribute to measure rather than a simple problem count. Another example shown in Fig. 1 concerns the effect of task complexity. For example, young children may be less able to execute complex interaction tasks than older ones. It is possible that the number and duration of the tasks should be smaller for younger children. We could then try to assess the effect of varying task complexity upon the problems reported or on the reliability of the method with the aim to determine the appropriate task length for children of different ages.

In the remaining of this section we present two studies addressing such questions, illustrating how they relate to the framework introduced.

8.1. Comparison of co-operative evaluation and co-discovery

This study concerned two usability testing methods with different procedures for prompting children to talk while interacting with an interface. The two different ways for encouraging children to verbalise what they do and think during interaction address their potential difficulty in verbalising their thoughts (see Fig. 1).

In the first method, which follows the set-up of a co-operative evaluation (Monk et al., 1993) participants work on representative tasks chosen by the usability specialist. One user and the usability specialist participate in an evaluation session. The user is encouraged to think aloud while working through the tasks and, if not, the usability specialist prompts the user to do so. The usability specialist allows the participant to make mistakes and uses unexpected behaviour and comments about the interface as symptoms of potential usability problems (Monk et al., 1993). The second set-up can be described as a co-discovery evaluation (Kemp and van Gelderen, 1996) where two users work collaboratively on tasks set by the usability specialist. An underlying assumption of co-discovery evaluation is that users will provide comments more naturally when they explore the interface together and therefore richer data will be gathered about expectations and opinions of the user.

To examine the consequences of these differences a usability test was conducted of a computer game for children called 'Junior Detectives'. It is a game for children aged 9–11 years, in which they have to solve problems (related to mathematics, language, geography and physics) as part of exploring a desert island on which the hero of the game has been stranded after a shipwreck. Since the main purpose of the study was to assess the two evaluation methods and not the computer game, only a part of the software was evaluated. Four pairs of children participated in the co-discovery sessions and four children participated in the cooperative evaluation sessions. The children were asked to work on four different tasks, covering various parts of the game. Each session took approximately 30 min. The boys and girls who participated in the study were aged 9–12. The usability evaluation sessions took place at an after-school care centre.

This case study concerned mostly the robustness of the two methods, i.e. the feasibility of applying these methods with children. To the best of our knowledge no literature described the application of these methods with children of this age group. In assessing
whether the method was applied appropriately, we also considered how the two methods compare with respect to how much children are influenced by the situation, for example by asking the facilitator for explanations? How much do the children influence each other in the co-discovery sessions? Do the children in the co-discovery method have more pleasure, because they are collaborating with a friend during the tasks? These questions relate also to the validity of the UTM output.

An overview of the data that was gathered to assess the influences of the differences between the methods on the outcome is given in Table 1. Overall, the methods seem to perform comparably: roughly the same number of problems was found by the two methods. Based on the limited number of times they asked the facilitator a question or explicitly looked at the facilitator, we infer that the children’s behaviour was hardly influenced by the presence of the facilitator in the co-discovery sessions. Furthermore, although some dominant behaviour was observed, both in the form of copying answers and in telling other children what to do, the impression was that relatively little information

<table>
<thead>
<tr>
<th>Measures</th>
<th>Co-discovery median [range]</th>
<th>Cooperative median [range]</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of usability problems that were uncovered</td>
<td>7.5 [6,9]</td>
<td>7[5,9]</td>
</tr>
<tr>
<td>The number of times the children ask for explanations of the facilitator</td>
<td>1 [1]</td>
<td>2.5 [2,4]</td>
</tr>
<tr>
<td>The number of times the children explicitly look at the facilitator</td>
<td>1.5 [1,2]</td>
<td>2 [1,3]</td>
</tr>
<tr>
<td>The number of times that the children copy the answer of the other child</td>
<td>2.5 [0,3] 0–8% of all questions</td>
<td>Not applicable</td>
</tr>
<tr>
<td>The number of times that one child takes action without discussing it first with other child, plus the number of times that one child tells the other child what to do, without providing an opportunity for another opinion</td>
<td>3.5 [3,5] 4–7% of all actions</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Number of expressions of frustration</td>
<td>3.5 [1,5] approximately 2% of all actions</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Expression of pleasure about participating in the session</td>
<td>2.5 [2,4]</td>
<td>0.5, 0.6 [0,1]</td>
</tr>
</tbody>
</table>
was lost. Finally, children in the co-discovery sessions appeared to enjoy more participating in the evaluation.

Based on this exploratory study it seems that the same kinds of insights will be found with these two methods. Of course, a more extensive study is needed to determine whether these findings hold for larger sets of subjects and different evaluators. It would also be important for a subsequent comparison study to choose a product or a set of tasks that has a larger number of usability problems, to be better able to determine the difference between two methods. When determining cost/benefits it depends on whether you count the number of problems found per subject, which is unfavourable for co-discovery, or on session-time, which is similar for co-discovery and cooperative evaluation in this study.

When choosing an evaluation method based on a combination of number of problems found and user-friendliness for the participants the co-discovery method is the best choice, based on this limited data. Especially since the possible drawback of the method of bias because of dominant children only occurs to a very limited extent. However, when counting the number problems found per subjects, co-discovery scores worse than cooperative evaluation, because it uses two children per session, versus only one child for cooperative evaluation.

8.2. Comparison of verbal protocol, interview and post-task questionnaire

The performance indicator adopted for the second study is the efficiency of the method, measured by the average number of problems found per child participant. We counted as problems the answers the children gave to the following questions: (a) Did you need help solving the task? (b) What, if anything, happened that you did not expect or want? Everything the children could not do without help was counted as a problem. Clearly this approach has drawbacks. We cannot be sure that all answers children gave should be really considered as usability problems or that all children consider them such. Apart from obvious references to the same problem and same interactive element (screen or object) we did not apply any theoretically founded scheme to classify usability problems. Our arguably simplistic approach has its advantages: it is independent of a possibly ephemeral understanding of what constitutes a usability problem; rather, it measures the efficiency of the method in generating problem reports, which is a rough but stable measure of comparison.

The usability testing methods varied with respect to the procedure by which children report usability problems. Three conditions were studied that can be described as concurrent verbal protocol, structured post-task interview and post-task questionnaire. In the verbal protocol condition children had been asked to report problems as they occurred. A printed sheet was kept at hand’s reach reminding them of what things to report (the two questions above) and if they were silent the facilitator would remind them to talk. In the interview condition, the facilitator would go through a structured questionnaire after every task that the children completed. In the questionnaire condition, the children would answer the same questions on paper. Children would then have to express their agreement with statements using a variant of Likert scales, proposed by (Hanna et al., 1997). These three reporting procedures are quite standard and together with their variants (e.g. retrospective verbal protocol, diaries) cover a large proportion of usability testing practice.
The characteristics of children hypothesised to have an effect on the number of problems reported were their verbalisation and the extroversion (see Fig. 1). Other variables that were studied were age and gender. The verbal protocol and the post-task interviews put some requirements on children to express themselves verbally, so these characteristics are arguably very relevant to the performance of children as test participants. Age and gender were not discounted, since the ability to spot and report problems could be reasonably argued to relate to both.

The three problem reporting procedures can be ranked according to the degree of active language skills required: think aloud > interview > questionnaire. Our hypothesis was that children that are better at verbalising and that are extrovert, would report relatively more problems when the usability testing method requires higher level of active language skills. On the other hand, children with relatively lower verbalisation and extroversion capabilities were expected to perform relatively better when the testing method would not require them to verbalise much. The outcome of the experiment is summarised below. For a more detailed exposition and an extensive discussion on its results, readers may consult Donker and Markopoulos (2002).

The verbal competence of the children was measured using four parts of the Wechsler Intelligence Scale for Children (WISC-R) (see De Bruyn et al., 1974). The parts chosen measure information, understanding, similarities and vocabulary. These sub-tests test various capabilities, one of which is productive language proficiency. The extroversion of the children was measured by letting them complete the extroversion items of the Amsterdamse biografische vragenlijst voor kinderen (ABV-K: the Amsterdam biographical questionnaire for children), see Wilde and Van Dijl (1967).

As far as possible tester intervention, tasks, external conditions, information given to the children were kept constant. In all cases a single child was involved with a single facilitator. Forty-five children of five different grades in three schools were asked to participate. The children were aged 8–14, with a mean age of 10 years and 2 months. Seven children were later excluded from the study. One was ill and with six there were technical difficulties to such an extent that the data became useless. The teacher introduced the experimenter to the class and nine children from each class were selected randomly. These children were first asked to complete the four sub-tests of the WISC-R and the extroversion items of the ABV-K. After that, they were randomly assigned to one of the

Table 2
Summary of Stepwise Regression Analysis for variables predicting the number for problems children found in the product (N = 38)

<table>
<thead>
<tr>
<th>Model</th>
<th>The number of problems found</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td>1</td>
<td>Constant</td>
</tr>
<tr>
<td></td>
<td>Think aloud</td>
</tr>
<tr>
<td>2</td>
<td>Constant</td>
</tr>
<tr>
<td></td>
<td>Think aloud</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
</tr>
</tbody>
</table>

Note. $R^2 = 0.427$ for Model 1; $\Delta R^2 = 0.106$ for Model 2 (p < 0.01). *p < 0.01.
three UTMs and they were asked to evaluate the program individually. To counteract a possible social desirability bias we stressed that we wanted to find as much awkward and illogical things in the program as possible.

A linear regression analysis was performed. The results are shown in Table 2. As it turned out our hypothesis was not verified. The number of problems found seemed to be determined by the testing method and the gender. Verbalisation and extroversion skills did not correlate significantly to the number of problems found. Note that this study was not set-up to study gender differences and a further measure of subjective satisfaction would have been necessary as a control to explain why girls report more problems than boys. While the original hypothesis was disconfirmed placing the study within the framework of Section 2 helps derive useful conclusions. Verbal protocol seems the most promising, suggesting that its variants would be interesting subjects of more detailed investigation. Our results are consistent with those found for adult users by Ohnemus and Beers (1993) and Henderson et al. (1995).

9. Conclusion and discussion

This paper has proposed a way to structure the problem space of comparing usability test methods for children. This was done by describing assessments of UTMs along three dimensions: the components that constitute the method, the measures for assessing a method and the special attributes of children as test participants. We contend that many such comparisons in the literature could be positioned along these first two dimensions. However, a closer look shows that perhaps the dimension we use to describe children’s characteristics is one that should describe characteristics of test participants in general. Arguably, comparative studies in the future should be sensitive to the fact that target user groups are very diverse and UTMs, like the products they aim to evaluate, need to be tailored for their target user groups.

In the paper we presented two studies that tried to answer some of the issues arising in the context of our framework. Some of our expectations were not verified. The empirical studies seem to open up more questions than they answer. However we hope that we are already shedding some light into an interesting and highly relevant sub-field of HCI.

Future work will try to extend these results by addressing several of the limitations of our work so far:

- **Typicality of post-hoc testing.** The contextual setting of a development environment is different than the experimental setting used here: usability evaluators know that their recommendations will incur costs; also problems are easier to find (in our case the product has probably gone through many stages of user testing).
- **What to report.** In order to provide useful results some researchers argue in favour of reporting process data (e.g. Jacobsen, 1999). Journals of usability tests with children could provide a useful resource for studying usability testing processes.
- **Definition of usability.** In the context of designing for children a different definition of usability might be applied for certain products. Since some products for leisure purposes are less task-oriented, a usability definition containing aspects of fun or
pleasure might have to be developed and applied. Although this issue has lately received more attention (e.g. Green and Jordan, 2002), no agreed upon extended definition of usability has been developed yet. At present we are conducting various studies to determine appropriate evaluation criteria for interactive products for children, as for example how to test not only for how usable a product is found but, also, how much fun children experience while interacting with it.

- **Typicality of product.** A significant problem is the generalisation we can make from a specific product. Can we define classes of products to choose representative products? Factors that might distinguish classes might be, use in the home or on the move, task-oriented products or play and fun oriented products, or individual or collaborative products.

- **Characteristics of children.** It is important to gain insights into how various characteristics of children can influence the applicability of usability testing methods. The list provided in this paper is by no means complete. We are conducting exploratory studies to determine the influence of characteristics of children on the suitability of evaluation methods, e.g. the extent to which children can think aloud and whether children of various age groups can fulfil more complex tasks.

The advantage of having a framework for comparing UTMs is that it helps us assess what we know versus what we still do not know, to relate research contributions and piecemeal research results that would otherwise remain fragmentary pieces of knowledge. In the long term, this framework can provide a basis for conducting a meta-analysis of comparative studies of usability testing methods for children, contributing to the cumulative development of knowledge in this area.

Acknowledgements

A. Donker and D. Van Breukelen are thanked for their effort during the case studies reported. We thank M. Kersten-Tsikalkina for the fruitful discussions during the writing of this paper.

References


