Evaluating fun and usability in computer games with children

Citation for published version (APA):

DOI:
10.6100/IR602065

Document status and date:
Published: 01/01/2006

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl

providing details and we will investigate your claim.
Evaluating fun and usability in computer games with children

Proefschrift

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een commissie aangewezen door het College voor Promoties in het openbaar te verdedigen op donderdag 9 februari 2006 om 16.00 uur

door

Wolmet Barendregt

geboren te Purmerend
Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. D.G. Bouwhuis
en
prof.dr. H. de Ridder

Copromotor:
dr.ir. M.M. Bekker
Contents

1 Including children in the design process ......................................................1
  1.1 Computers at an early age..............................................................................2
  1.2 Involving children in User-Centered Design....................................................2
  1.3 Formative and summative evaluation............................................................4
  1.4 Intended age group .......................................................................................4
    1.4.1 Hanna, Risden & Alexander .................................................................5
    1.4.2 Valkenburg ..............................................................................................6
    1.4.3 Acuff and Reiher ....................................................................................6
    1.4.4 Piaget .....................................................................................................6
    1.4.5 Age group for this thesis .......................................................................7
  1.5 Usability and fun ..........................................................................................8
  1.6 Game genres ...............................................................................................9
  1.7 Research questions and thesis outline .........................................................9
    1.7.1 Chapter 2: Adapting the evaluation process for children and computer games..10
    1.7.2 Chapter 3: Detecting problems in games by analysing video data ..........11
    1.7.3 Chapter 4: The influence of practice on problem detection ................11
    1.7.4 Chapter 5: Personality characteristics as predictors of participant effectiveness12
    1.7.5 Chapter 6: The Problem Identification Picture Cards Method ..............12
    1.7.6 Chapter 7: Testing with children in practice ........................................13
    1.7.7 Chapter 8: Discussion and conclusions ..............................................13

2 Adapting the evaluation process for children and computer games..............15
  2.1 Ethics concerning user testing with children ...............................................16
    2.1.1 Consent .................................................................................................16
    2.1.2 Deception ..............................................................................................16
    2.1.3 Debriefing ..............................................................................................16
    2.1.4 Withdrawal from the investigation .....................................................16
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.5</td>
<td>Confidentiality</td>
<td>17</td>
</tr>
<tr>
<td>2.1.6</td>
<td>Protection of participants</td>
<td>17</td>
</tr>
<tr>
<td>2.1.7</td>
<td>Privacy during observational research</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td>Evaluating work and play</td>
<td>17</td>
</tr>
<tr>
<td>2.3</td>
<td>Does providing tasks lead to natural playing behaviour?</td>
<td>19</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Experimental (pilot) study: The influence of tasks</td>
<td>19</td>
</tr>
<tr>
<td>2.4</td>
<td>Verbalisation techniques with children</td>
<td>21</td>
</tr>
<tr>
<td>2.5</td>
<td>Giving help during evaluations</td>
<td>23</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Providing the right help</td>
<td>25</td>
</tr>
<tr>
<td>2.5.2</td>
<td>The adapted Interaction Cycle</td>
<td>27</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Risks of providing help in the wrong phase</td>
<td>29</td>
</tr>
<tr>
<td>2.6</td>
<td>Conclusion</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Detecting problems in games by analysing video data</td>
<td>33</td>
</tr>
<tr>
<td>3.1</td>
<td>The evaluator effect in analysing video material</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>Analysis of video data</td>
<td>34</td>
</tr>
<tr>
<td>3.3</td>
<td>The coding scheme</td>
<td>38</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Nature of games</td>
<td>40</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Fun problems</td>
<td>41</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Specific behaviour of children playing games</td>
<td>43</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Adaptations to eliminate overlap between indications</td>
<td>45</td>
</tr>
<tr>
<td>3.3.5</td>
<td>The proposed coding scheme</td>
<td>45</td>
</tr>
<tr>
<td>3.4</td>
<td>Measuring the reliability of the coding scheme</td>
<td>47</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Analysis 1: Determining any-two agreements</td>
<td>49</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Analysis 2: Determining Cohen’s kappa for fixed observation points</td>
<td>51</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Analysis 3: The influence of training</td>
<td>52</td>
</tr>
<tr>
<td>3.5</td>
<td>Discussion</td>
<td>53</td>
</tr>
<tr>
<td>3.6</td>
<td>Conclusions</td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td>The influence of practice on problem detection</td>
<td>57</td>
</tr>
<tr>
<td>4.1</td>
<td>Practice before the evaluation</td>
<td>58</td>
</tr>
<tr>
<td>4.2</td>
<td>Kinds of knowledge development</td>
<td>59</td>
</tr>
<tr>
<td>4.3</td>
<td>Theory of human errors</td>
<td>60</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Usability problems</td>
<td>61</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Inefficiencies</td>
<td>64</td>
</tr>
<tr>
<td>4.4</td>
<td>Intrinsic motivation in computer games</td>
<td>64</td>
</tr>
<tr>
<td>4.5</td>
<td>Experiment: Research questions and hypotheses</td>
<td>65</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Usability problems</td>
<td>65</td>
</tr>
</tbody>
</table>
4.5.2 Knowledge inefficiencies ...................................................................................... 66
4.5.3 Fun problems ....................................................................................................... 67
4.5.4 Severity measures for problems .......................................................................... 68
4.5.5 Efficiency, effectiveness, and satisfaction ............................................................ 69
4.5.6 Previous experience ............................................................................................. 69
4.6 Method ..................................................................................................................... 70
4.6.1 Participants .......................................................................................................... 70
4.6.2 Test material ......................................................................................................... 71
4.6.3 Procedure ............................................................................................................. 72
4.7 Analysis of the test sessions ..................................................................................... 73
4.7.1 Introduction ......................................................................................................... 73
4.7.2 The analysis procedure for the last test session ..................................................... 74
4.8 Reliability analyses ............................................................................................... 74
4.8.1 Reliability of the problem analysis of the two sessions ......................................... 74
4.8.2 Reliability of the problem classification ............................................................... 74
4.9 Results ...................................................................................................................... 75
4.9.1 Types of problems in both test sessions ............................................................... 75
4.9.2 Hypotheses problem types .................................................................................. 76
4.9.3 Problem severity ................................................................................................. 79
4.9.4 Effectiveness, efficiency, and satisfaction ............................................................ 83
4.9.5 Experience .......................................................................................................... 83
4.10 Discussion ............................................................................................................... 84
4.10.1 Considerations about the unconfirmed hypotheses ............................................. 84
4.10.2 Limitations ......................................................................................................... 85
4.10.3 What else can be learned from a comparison of both test sessions? ................. 86
4.11 Conclusions ............................................................................................................ 87

5 Personality characteristics as predictors of participant effectiveness ...... 89
5.1 Differences between children .................................................................................. 90
5.2 User test outcome and personality characteristics .................................................. 91
5.2.1 User test outcome ............................................................................................... 91
5.2.2 Personality characteristics .................................................................................. 92
5.3 Experiment: Hypotheses ......................................................................................... 95
5.4 Exploring other predicting factors .......................................................................... 96
5.5 Representativeness of problems after selection ....................................................... 97
5.6 Method ..................................................................................................................... 97
5.6.1 Participants .......................................................................................................... 97
5.6.2 Test material ....................................................................................................... 98
5.6.3 Procedure .............................................................................................................98
5.7 Analysis ...................................................................................................................98
5.8 Determining the percentage of verbally indicated problems.................................99
5.9 Determining representativeness ..............................................................................99
5.10 Results .................................................................................................................101
  5.10.1 Hypotheses ....................................................................................................101
  5.10.2 Exploring other factors ..................................................................................101
  5.10.3 Representativeness .........................................................................................106
  5.10.4 Example of the effects of selection ................................................................108
5.11 Discussion ............................................................................................................110
  5.11.1 Related research ............................................................................................110
  5.11.2 Generalisability .............................................................................................111
  5.11.3 Practical advice ...............................................................................................112
5.12 Conclusions .........................................................................................................112

6 The Problem Identification Picture Cards method ..............................................115
  6.1 Thinking-aloud with children .............................................................................116
  6.2 Development of the method ..............................................................................116
  6.3 Choosing the pictures .........................................................................................118
  6.4 The Problem Identification Picture Cards method .............................................121
  6.5 Experiment: Evaluation of the PIPC method .....................................................121
    6.5.1 Hypothesis 1 ................................................................................................122
    6.5.2 Hypothesis 2 ................................................................................................122
    6.5.3 Hypothesis 3 ................................................................................................122
  6.6 Method ................................................................................................................123
    6.6.1 Test participants ............................................................................................123
    6.6.2 Experimental set-up .....................................................................................123
    6.6.3 Test material ................................................................................................124
    6.6.4 Procedure ....................................................................................................124
  6.7 Analysis ...............................................................................................................125
  6.8 Results ................................................................................................................126
  6.9 Discussion ............................................................................................................130
    6.9.1 Games ..........................................................................................................130
    6.9.2 Procedure ....................................................................................................130
    6.9.3 Possible improvements ................................................................................131
    6.9.4 An unpredicted benefit of the PIPC method ................................................132
  6.10 Conclusions .......................................................................................................132

7 Testing with children in practice ........................................................................135
7.1 Guidelines to conduct usability tests with children ............................................. 136
7.2 Who will be the test participants? ........................................................................ 136
  7.2.1 Existing guidelines ......................................................................................... 136
  7.2.2 The number of children .................................................................................. 137
  7.2.3 Selecting children .......................................................................................... 138
  7.2.4 Children with disabilities ............................................................................... 139
7.3 Where will we test? ............................................................................................. 139
  7.3.1 Existing guidelines ......................................................................................... 139
  7.3.2 Testing at schools ........................................................................................... 139
7.4 What will we ask the test participants to do? ..................................................... 140
  7.4.1 Existing guidelines ......................................................................................... 140
  7.4.2 Tasks or free play ........................................................................................... 140
  7.4.3 Making the children stop ............................................................................... 140
7.5 How will we organise the test session(s)? ......................................................... 141
  7.5.1 Existing guidelines ......................................................................................... 141
  7.5.2 Allowing children to practice before the actual test ....................................... 142
  7.5.3 Using a script to explain the purpose of the test .......................................... 143
  7.5.4 Using a script during the test ......................................................................... 143
  7.5.5 Giving help .................................................................................................... 144
  7.5.6 Thinking aloud ............................................................................................... 145
  7.5.7 Using the Problem Identification Picture Cards method ................................ 145
  7.5.8 Questionnaires .............................................................................................. 146
  7.5.9 Gifts ................................................................................................................ 146
7.6 How will we organise the data analysis? ............................................................ 147
  7.6.1 Existing guidelines ......................................................................................... 147
  7.6.2 Using a coding scheme .................................................................................. 147
  7.6.3 Number of evaluators .................................................................................... 148
7.7 Discussion and conclusions ............................................................................. 148

8 Discussion and conclusions .............................................................................. 151
  8.1 Recapitulation .................................................................................................... 152
  8.2 Limitation of the research: problem fixing and redesign ................................. 156
  8.3 Future work ....................................................................................................... 157
    8.3.1 Why are we testing? ....................................................................................... 157
    8.3.2 Who will be the test participants? ................................................................. 157
    8.3.3 Where will we test? ....................................................................................... 158
    8.3.4 What will we ask the test participants to do? .............................................. 158
    8.3.5 How will we organise the test session(s)? .................................................. 158
8.3.6 How will we organise the data analysis? ................................................................. 159
8.3.7 How will we organise the reporting? .................................................................. 159

Bibliography ................................................................................................... 161
Summary ........................................................................................................ 173
Samenvatting .................................................................................................. 175
Curriculum Vitae ............................................................................................ 179
Dankwoord ..................................................................................................... 181
Appendix 1 Previous experience questionnaire and permission form ............ 185
Appendix 2 Preference for game and method questionnaire ......................... 187
Appendix 3 Consent form ............................................................................... 189
1 Including children in the design process

This first chapter presents the general research topic of this thesis: How can we increase the amount of relevant information yielded by observational evaluations with children between five and seven years old, about usability and fun problems that should be fixed in order to improve games? The chapter starts with a discussion of the need for observational evaluations of computer games with children. Furthermore, it is explained that this thesis will focus on formative evaluations in which the purpose of the evaluation is to detect and describe problems instead of determining general quality measures like effectiveness, efficiency, and satisfaction. Based on an overview of the literature on the different stages of child development and age groups for several areas of interest, the selection for the specific age group of children between five and seven years old is substantiated. Based on theories of user’s needs, it is explained why in this thesis it is assumed that the aim of a formative evaluation of a computer game is to get information about both fun problems and usability problems, instead of only about usability or fun problems. Furthermore, the game genre that will be considered in this thesis is described. At the end of this chapter the different research topics are presented in order to answer the main research question. The research topics include a number of innovations in procedure, participant selection, and data analysis.
1.1 Computers at an early age

When computers were first developed they used to be the domain of experts, like engineers and programmers, putting a lot of effort in trying to operate them (Grudin, 1990). Nowadays, computers are used by almost everyone and also children get exposed to computers and technology at an increasingly early age. For example, there are educational CD-ROMs for children as young as 18 to 36 months, and special keyboards have been developed for babies and small children (e.g. by Génération5 (2005), Berchet (2005), and Ergocube (2005)).

![Example of a keyboard for children](image)

**Figure 1-1 Example of a keyboard for children (with permission of Ergocube).**

For most children the first contact with the computer is through some sort of (educational) game, and children play computer games very often. In The Netherlands in 2005 61% of the children under 15 play computer games every day (ANP, 2005). It is therefore important that computer games for children are well-designed for the intended age-group.

1.2 Involving children in User-Centered Design

One of the most commonly used design philosophies to create high quality products for users is the User-Centred Design (UCD) approach (Norman and Draper, 1986, Rubin, 1994, Nielsen, 1993). UCD refers to the philosophy that the intended user of a product should always be in the centre of the design process throughout all phases of the design. Druin (1999) gives a classification of the different roles children can play during the design process; children can be users, testers, informants, or design partners. Although the levels of engagement are different for the different roles, they all include evaluations with child participants as evaluators. This means that that products should be evaluated by having children use an actual implementation of the product in some form in a representative way.
There are three types of evaluation methods: observational evaluation methods, analytical evaluation methods and inquiry methods.

Evaluation methods that collect data by observing users’ experiences with a product are called observational evaluation methods. These methods have a user performance view (Rauterberg, 1992). Some types of observational evaluation methods are the usability test (Lewis, 1982), the user performance test (Nielsen, 1993), and cooperative evaluation (Wright and Monk, 1991). Methods that do not collect data from users’ experiences but rely on the opinion of experts are called inspection or analytical evaluation methods. These methods have a product oriented view (Rauterberg, 1992). Examples of analytical evaluation methods are the Heuristic Evaluation (Nielsen and Molich, 1990) and the Cognitive Walkthrough (Lewis et al., 1990). Inquiry methods focus on users’ likes and dislikes, needs, and understanding of a product by asking users to answer questions verbally or in written form. These methods have a user oriented view (Rauterberg, 1992). Inquiry methods tend to identify broad usability problems or opinions about a product as a whole. Examples of inquiry methods are User Satisfaction Questionnaires (Reiterer and Oppermann, 1993) and Focus Groups (Zirkler and Ballman, 1994).

While all these methods are directed at adult users, it should be realised that young children are less able to read, verbalise, concentrate, and perform abstract logical thinking than adults (Markopoulos and Bekker, 2002). Inquiry methods which rely on these skills are therefore not very well applicable with young children. Furthermore, because these methods tend to identify broad usability problems or opinions about a product as a whole they are less suitable to acquire knowledge about specific problems for the purpose of improving a product. Many products for children are still analytically evaluated by adult experts (Buckleitner, 1999) instead of observationally because observational evaluation methods that work with adults can not always be applied with children (Hanna et al., 1997). However, “it is not easy for an adult to step into a child’s world” (Druin, 1999), and sometimes children show behaviour with a game that is very difficult for adult experts to predict. For example, in the computer game ‘Rainbow, the most beautiful fish in the ocean’ (MediaMix, 2002b) young children often try to use speech input with a game due to the type of instruction the character gives. When the character says: ‘Would you help me count the scales that I still have to find?’ many children spontaneously respond by starting to count aloud instead of clicking the
scales. While most adults know this kind of input is probably not appropriate, young children often think that the computer can hear them and do not attempt to click anything. Examples like this show that actually involving young children in the observational evaluation of computer games is highly desirable. However, the difference between the capabilities of young children and adults may require adaptations to the applied evaluation methods. This thesis focuses on how the amount of relevant information, yielded by observational evaluation methods during the evaluation of computer games with young children, can be increased.

1.3 Formative and summative evaluation

Often, observational usability evaluations are performed to determine quantitative measures like efficiency, effectiveness, and satisfaction (ISO, 1998). These measures can be used to compare or assess the level of usability of an entire product. Evaluations in order to determine these measures are called summative evaluations (Hartson et al., 2001). However, another common goal is to identify as many aspects of a product that cause users trouble (Hertzum and Jacobsen, 2001) for the purpose of improving the product by fixing these problems. This type of evaluation is often called formative evaluation (Barnum, 2002, (Hartson et al., 2001). By involving children in formative evaluations of computer games it is possible to improve the games based on their input. In this thesis we therefore aim to increase the amount of information from children for downstream causal analysis of the problems that should be fixed to improve the quality of a game.

In the remainder of this thesis the term user test refers to this latter practice of detecting problems and getting information about these problems in order to fix them. Thus, the term user test here does not refer directly to assessing the quantitative measures efficiency, effectiveness, and satisfaction. However, solving the detected interaction problems in a product should, in turn, lead to increases in efficiency, effectiveness, and satisfaction.

1.4 Intended age group

In growing up children’s capabilities to read, verbalise, concentrate, and perform abstract logical thinking increase (Markopoulos and Bekker, 2002). Therefore, the research of observational evaluation methods for computer games with children needs a clear demarcation of the intended age group.
First of all, the children in the targeted age group should be clearly different from adults. Furthermore, the children in the chosen age group should be able to participate in an observational evaluation. Therefore as a rough selection we restrict ourselves to children between four and eight years old. In this section some of the literature from different perspectives of the characteristics of children between four and eight years old is described. Based on this literature a more specific age group is defined at the end of the section.

Hanna, Risden, and Alexander’s categorisation of age groups for usability testing (1997) is related most closely to the topic of this thesis. Furthermore, the age group characterisations from a marketing (Acuff and Reiher, 1997) and consumer (Valkenburg, 1999) point of view are given because they relate to what children like to buy or use. Finally, the characterisation of Piaget (1970) is given as his theory of cognitive development has had a great influence on all modern developmental psychologists.

1.4.1 Hanna, Risden & Alexander

In their guidelines for usability testing with children Hanna, Risden, and Alexander (1997) discern three age groups: 2 to 5 years, 6 to 10 years, and 11 to 14 years. For usability evaluations they state that children younger than 2 ½ years are not proficient enough with standard input devices to provide useful data, and children older than 14 behave so much like adults that they should be treated accordingly. Children between four and eight years old are part of the youngest and the middle age group. Some more detailed characterisations of these age groups according to Hanna, Risden, and Alexander (1997) are:

- **2 to 5 years:** Preschool children’s attention span is rather low. Furthermore, their motivation to please adults and their ability to adjust to unfamiliar surroundings and new people may change from one moment to the next, making the course of the test session unpredictable. Therefore, this age group requires the most extensive adaptations of user testing.

- **6 to 10 years:** Children in this age group are used to go to school. Therefore they can perform tasks successfully and follow directions from adults. They will answer questions and try out new things, and they are not very concerned about being observed when they play with the computer. The youngest children of this age group (6 and 7 years old) may be a little shy or inarticulate when talking about the computer.
1.4.2 Valkenburg

Valkenburg (1999) discerns four phases in the transition from child to consumer: 0-2 years, 2-5 years, 5-8 years, and 8-12 years.

- 2-5 years: In this period children start to understand cause-and-effect relations. In this stage the fantasy of children develops rapidly and they are not always able to see the difference between reality and fantasy. Children in this age group tend to focus all their attention to one eye-catching attribute of an object or information.

- 5-8 years: The attention span of children in this age group is much larger than that of children in the earlier stage. Three-year old children can only keep their attention to a single activity for about thirty minutes and are still often distracted, whereas five-year old children can sometimes concentrate on a favourite activity for more than an hour.

1.4.3 Acuff and Reiher

In their book about marketing for children Acuff and Reiher (1997) discern five age groups: 0-2 years, 3-7 years, 8-12 years, 13-15 years, and 16-19 years.

- 3-7 years: Children in this age group are the most playful of all age groups. Imagination, fantasy, surprise, and pretend play, play an important role in their lives. They are not yet able to retrace their thinking-process back to where it started, therefore they cannot give logical explanations for their actions. As a result children in this age group are rather impulsive and reactive in their approach to the world, rather than logical, reasonable, or reflective.

- 8-12 years: Children in this age group are very sensitive to peer acceptance and opinion. They often feel the need to turn away from childish concepts. Around 7 or 8 years old children begin to be able to handle simple abstractions, and think more rationally and logically. They also start to pay attention to details and appreciate them.

1.4.4 Piaget

Piaget (1970) proposed a theory of cognitive development in childhood in which he discerned four main stages of child development: 0-2 years (sensorimotor stage), 2-7 years (preoperational stage), 7-11 years (concrete operational stage), and 11 years and older (formal operational stage). Further, Piaget believed that cognitive performance in children is directly associated with the cognitive development stage they are in. So, if a child were in the
preoperational stage, he/she would not be able to successfully master tasks of a concrete operational stage child. There have been many criticisms of Piaget’s theory (Lourenco and Machado, 1996). Most notably, developmental psychologists debate whether children actually go invariably through these four stages in the way that Piaget proposed, and further that not all children reach the formal operation stage. Despite these criticisms, Piaget has had a major influence on all modern developmental psychologists and the four major stages of development are still often used.

- 2–7 years (Preoperational stage): Children in this stage acquire representational skills in the area of mental imagery, and especially language. They are very self-oriented, and have an egocentric view; that is, preoperational children can use these representational skills only to view the world from their own perspective.

- 7–11 years (Concrete operational stage): During the Concrete Operations Stage children develop logic and are able to use rules and units. At the same time, their thoughts remain closely linked to the present, involving real objects.

1.4.5 Age group for this thesis

To make the differences between the youngest and oldest children in the research smaller, thus creating an age group with children of very similar capabilities, it was decided to narrow the age group of four to eight years old to children between five and seven years old. This means that the youngest children have just turned five years old while the oldest children have not yet turned eight years old. In the remainder of this thesis the phrase “between five and seven years old” will be used to indicate this chosen age group.

Children between five and seven years old form a very interesting group for this thesis because according to Hanna et al. (1997) they may require extensive adaptations of traditional usability testing. Furthermore, many children in this age group already play computer games quite often, making it important to include them in the evaluation of games.

On the one hand children in this age group cannot yet read very well, they are impulsive and reactive in their approach to the world, rather than logical, reasonable, or reflective, and they may have difficulties verbalizing their thoughts when talking about the computer. On the other hand they are suitable participants for the User-Centred Design practice because they can concentrate long enough to perform an evaluation in a usability lab or at school.
1.5 Usability and fun

Formative evaluations of work-related products are usually aimed at finding usability problems. When a usability problem is encountered it means that the user is not able to reach his/her goal in an efficient, effective, or satisfactory way. However, in computer games the key factors for success are pleasure and fun (Pagulayan et al., 2003). When at a certain point a game is not pleasurable or fun to play, when it is boring, too difficult, or too easy, the user encounters a fun problem. Beside usability problems fun problems form another category of problems that can occur in a computer game. These problems should also be detected and described in a formative evaluation because they are just as important to fix as usability problems, if not even more important. It is however not really possible to look only for usability or fun problems because usability and fun are linked closely together. Jordan (1999) proposed a hierarchy of user needs, in turn based on Maslow’s (1970) hierarchy. This hierarchy can also be applied to computer games.

Figure 1-2 Jordan’s (1999) hierarchy of player needs applied to a computer game.

The idea of this hierarchy is that when users have satisfied the needs lower down the hierarchy they want to satisfy the needs higher up. Satisfying the needs on a lower level is also a prerequisite for satisfying the needs higher up.

Applying this hierarchy to a computer game means that, most importantly, the intended goals of the game should be reachable. This is the functional level. For example, when the goal of a game is to rescue the princess, the game should make sure that the princess can be rescued. On the usability level, the ways to play the game should be clear, and controls should
not provoke discomfort. Finally, on the pleasure level, the game needs to be fun, by providing interesting challenges and a fantasy that is appealing to the user. This is the reason why it is not possible in an evaluation of a game to solely detect either usability or fun problems. Both types of problems will occur during the test and will influence each other. Furthermore, the way the user responds to both types of problems may be very similar. Of course, to improve the game after the problems have been detected it may be necessary to distinguish between these different types of problems because they may require different solutions. In this thesis the detection and description of both fun problems and usability problems is considered a necessary result of an observational evaluation. The detection of functionality problems is considered to be a debugging process which should be performed by experts before the users asked to perform an evaluation.

1.6 Game genres

A very common game genre for children between five and seven years old is the adventure game. These games often have roughly the same structure of several sub games that have to be played to reach a certain overall goal. There is usually no fixed path to play the sub games, and children have some limited choice which sub games they want to play in which order. Most of the sub games form a step towards a main goal. Among these sub games there are e.g. motor-skill games, logical games, and musical games and they often also have some educational value (Donker, 2005). The games can usually be finished in one or two hours of playing. There are some other genres of games available for children in the chosen age group for example simulation games (an example is Roller Coaster Tycoon (Atari, 1999) which is actually meant for children of eight and up but during the tests some children said they already played it), racing games e.g. RedCat SuperKarts (Davilex, 2000), and action games like mazes e.g. Freddi Fish, Silly Maze (Transposia, 2002), but this thesis will focus on adventure games for the home-market.

1.7 Research questions and thesis outline

The main research question of this thesis is: How can we increase the amount of relevant information yielded by observational evaluations with children between five and seven years old, about usability and fun problems that should be fixed in order to improve games? To answer the main research question, a number of innovations in procedure, participant selection, and data analysis will be investigated in this thesis. These innovations will extend
the research field of usability testing with children by specifically aiming at the evaluation of computer games, and they will extend the depth of the obtained information in order to improve the games by selecting suitable participants and proposing an addition to thinking-aloud. All innovations can be classified according to a framework for user testing: why are we testing, who will be test participants, where will we test, what will we ask the test participants to do, and how will we organise the test session(s), the data analysis, and the reporting. The following sections give an overview of the topics discussed in each of the chapters of this thesis, the related question(s) in the framework, and the sub questions answered in each chapter.

1.7.1 Chapter 2: Adapting the evaluation process for children and computer games

To protect the children who will participate in the experiments it is necessary to decide how an ethical approach will be guaranteed (Patel and Paulsen, 2002). The precautions taken in this research are discussed in section 2.1. This topic relates to how the test sessions will be organised.

In order to observe and get relevant information about problems in games it has to be decided whether children should play freely or receive some fixed tasks to perform. In section 2.3 it is discussed whether concrete tasks, like navigating to a particular part of the game, or playing a certain sub game should be given when evaluating both fun and usability in computer games. This research question relates to what we will ask the test participants to do.

When performing an evaluation in order to detect problems that need to be fixed it is necessary that the children only help to detect problems that they really experience. In section 2.4 it is discussed whether children should be prompted to think-aloud during the evaluation, as is normally done with adults. This topic relates to determining how the test sessions should be organised.

When children ask for help during an evaluation, as often happens, it can be a good indication of a problem for the evaluator who is analysing the test data. However it is necessary to decide how the facilitator, who will be conducting the tests with the children, should respond to these questions for help. Section 2.5 discusses how and when the facilitator should give help to the children. This relates to determining how the test sessions should be organised.
1.7.2 Chapter 3: Detecting problems in games by analysing video data

Many studies that count problems yielded by observational evaluation methods are unclear about how the evaluators, the people analysing the video data to create problem reports, should decide that there is a problem. For example, in Als et al. (2005) it was only stated that two researchers conducted the evaluations without stating whether the researchers used a predefined list of problem indicating behaviours. Such a list of behaviours is necessary to report about reliability of the evaluation. These problem-indicating behaviours form a coding scheme and several researchers have provided coding schemes for the detection of problems in products for adults (Vermeeren et al., 2002, Cockton and Lavery, 1999). In Chapter 3 the development and evaluation of a coding scheme to detect both fun and usability problems in computer games with young children is described. The subsequent chapters all use this coding scheme for detecting problems. Chapter 3 thus relates to determining how the data analysis should be organised.

1.7.3 Chapter 4: The influence of practice on problem detection

The usability and fun problems that are detected during an evaluation should be relevant problems. This means that the detected problems should have a negative influence on the usability and fun of the game, and fixing them will improve the overall quality of the game. Although user testing is often performed for first use, the differences between novice and expert use provide insights in what may happen to problems experienced by novice users when they get more experienced (Kjeldskov et al., 2005). It may also be the case that the problems detected in an evaluation with children who play a game for the first time are different from the problems that they experience after some practice, even after a short time. The results of an evaluation after practice may therefore merit other changes to the game than the results of an evaluation of first use. However, it has not been thoroughly investigated what kind of problems children encounter when they play a game for the first time and what kind of problems children encounter when they have practiced with a game. Furthermore, the problems that children encounter during first time may still be present after some practice but their severity may have changed, resulting in a different decision of which problems need to be fixed first. Finally, the standard usability metrics Efficiency, Effectiveness, and Satisfaction may change when children practice with a game. In Chapter 4 the effects of practice on the kinds of problems, the severity of problems and the overall
usability assessment of a game are investigated. Chapter 4 thus relates to determining how the test sessions will be organised.

1.7.4 Chapter 5: Personality characteristics as predictors of participant effectiveness

In practice, most evaluations are performed with a small number of participants because conducting and analysing test sessions is expensive. However, there are many differences between test participants, whether they are adults or children. In order to increase the amount of information that an evaluation yields about problems that need to be fixed it could be beneficial to select children as test participants who are able to detect many problems and talk about them. Other researchers have done this in a fairly rough manner. For example, for the evaluation of game concepts, Hanna et al. (2004) selected children who were characterised by their parents as not shy. However, it has not yet been investigated whether personality characteristics are actually good predictors of how well children and adults are able to find problems and talk about them. Furthermore, it has not yet been investigated whether the selection of participants results in finding representative problems for all users. In Chapter 5 an experiment is described to determine whether it is possible to predict which children will be able to help detect many problems and provide clear information by talking about these problems. Furthermore, the representativeness of the problems detected by a selection of children is investigated. Chapter 5 thus relates to determining who should be the test participants in order to increase the amount of relevant information.

1.7.5 Chapter 6: The Problem Identification Picture Cards Method

Although the effectiveness of an evaluation may be improved by selecting suitable participants as was proposed in Chapter 5, this is not always possible. For example, there may be only a few children available as test participants. In Chapter 6 a new evaluation method is developed and evaluated in order to help children express more usability and fun problems than standard thinking aloud. This method is called the Problem Identification Picture Cards (PIPC) method and it makes use of a box with picture cards symbolizing different types of problems that children can encounter while playing a game. These picture cards can be used to explain the purpose of the test and serve as a reminder during the test. Finally, the picture cards can be used by less articulate children to express problems clearly in a non-verbal way by putting a picture card in the box. The first sub question is: Does the
PIPC method help children to express more problems than thinking aloud? The second sub question is: Will children who use the PIPC method still verbalise the problems they encounter, or will they just replace their verbalisations with picture cards? The last sub question is: Do children find new method pleasurable to use in comparison with standard thinking aloud? The experiment to answer these sub questions is described in Chapter 6. Chapter 6 thus relates again to determining how the test sessions should be organised.

1.7.6 Chapter 7: Testing with children in practice

Hanna et al. (1997) provide the most commonly cited guidelines for usability testing with children. These guidelines are based on their experiences as usability engineers. However, no experimental data are given to support their guidelines. In Chapter 7 the advice of Hanna et al. (1997) is summarised shortly and the results of the experiments in this thesis are added to extend these guidelines for practitioners wishing to evaluate computer games with young children. Furthermore, some personal experiences in the labs and at schools are added. The advice will be organised according to the seven questions of the framework for conducting user tests.

1.7.7 Chapter 8: Discussion and conclusions

Chapter 8 recapitulates the research questions and conclusions of this thesis. Limitations and possible directions for further research based on the framework for conducting a user test are discussed in more detail.
2 Adapting the evaluation process for children and computer games

Evaluating computer games for children is different from the standard practices of evaluating other products for adults. First of all, children are a vulnerable group of participants to do research with. Therefore, it is necessary to decide how an appropriate ethical approach will be guaranteed in the experiments performed for this thesis. Furthermore, not only do games have different properties than other products like coffee machines or word-processing programs, but also children are different from adults. Therefore the common practice of giving tasks during an evaluation is discussed in relation to natural interaction behaviour with games. The assumption is that the assigned tasks interfere with the natural way of playing a game; a pilot experiment to substantiate this assumption is presented. Subsequently, the use of the thinking-aloud method with children is discussed. Arguments are provided for the guideline that children should be asked to verbalise as much as possible but they should not be prompted to think aloud when they forget to do so, or have difficulty doing it. Lastly, it is discussed why help was given to the children in the experiments presented in this thesis and it is suggested how facilitators should give help based on a common model of user-system interaction.

1 Part of this chapter is based on the following publication: (Barendregt et al., 2003)
2.1 Ethics concerning user testing with children

In order to perform the experiments in this thesis many children will participate in user tests. Special ethical considerations are involved in testing with children. To my knowledge there are no laws in The Netherlands for conducting user tests with children. However, the British Psychological Society (British Psychological Society, 2005) provides principles for conducting ethical research with human and child participants. These principles include getting consent, handling deception, debriefing the participants, offering the possibility to withdraw from the experiment, confidentiality, protection of participants, observational research, and giving advice. Each of these principles and how they are implemented in the experiments in this thesis will be discussed in the subsequent sections.

2.1.1 Consent

In the user tests performed for this thesis the children will get an explanation from the facilitator about the purpose of the test. One of the parents will have to sign the consent-form which is given in Appendix 3. The children (and their parents when they are present during the test) will receive a small gift of around €2.00 (≈ $2.00), but they will not be informed about this gift before the test takes place.

2.1.2 Deception

The children in the user tests performed for this thesis will not be deceived. They will always know that the purpose of the test is to get information from them about problems that need to be fixed in order to improve the tested game.

2.1.3 Debriefing

The nature of the test sessions in which children are asked to play a game is such that the children will not need any other information to complete their understanding of the research than what they already receive before the test session.

2.1.4 Withdrawal from the investigation

The children in the experiments will be told they can withdraw from the experiment at any time. The facilitator will remind them of their right to withdraw whenever they are reluctant to play any further before the end of the test session.
2.1.5 Confidentiality

The parents of the children participating in the experiments will be able to indicate how they want the data of their children to be used on the consent form (see Appendix 3). They can either choose to give permission for analysis of the data without the possibility of using the gathered video material in research-related presentations, or they can choose to make the video data available for research-related presentations.

For the experiment in Chapter 5 a personality questionnaire will have to be filled out by the parents about their children. The results of this questionnaire will be presented anonymously.

2.1.6 Protection of participants

Investigators have a primary responsibility to protect participants from physical and mental harm during the investigation. In research involving children, great caution should be exercised when discussing the results with parents, or caretakers, since evaluative statements may carry unintended weight.

The children in the experiments will not be exposed to any risk or harm which goes beyond what can be expected in ordinary life. Furthermore, the results of the personality questionnaires will not be discussed with the parents.

2.1.7 Privacy during observational research

All children will be informed that they will be observed by means of one or more cameras. They will be made aware of the positioning of the cameras and they will be allowed to look around in the observation room containing the video equipment when the experiment takes place in the lab. Furthermore, if assistants are present during the test the children will be introduced to the assistants before the test.

2.2 Evaluating work and play

When performing a usability test one can choose between presenting the test participants with fixed tasks and letting them explore the product by themselves without any fixed tasks. During the test the thinking-aloud technique is used to capture what the participants are thinking in order to detect problems with the product and get an indication of their causes.

An example taken from Carroll and Mack (1984) of what a test participant may say and do when trying to delete the underscore from the second of three words “will not change”. The
test participant is not locating the cursor in a way that identifies the special underscore character:

*What happened? The line [underscore line] didn't disappear, now what do I do? Was I over too many spaces? I could have been.

Let me try again. I can replace the other one anyway.*

[Presses character delete, deleting the “c” of the third word “change”. Not clear why she did this.]

*What’s happening?*

[Positions cursor immediately after the third word and ‘under’ the underscore instruction, so that the character delete operation correctly deletes the underscore from that word]

*It worked on that one. I don’t understand*

[Here is the problem: Underscore is represented by a special character symbol. To delete, the cursor must be positioned ‘under’ this symbol and then deleted as any other character]

[Presses word delete key, and deletes the third and fourth words]

*I just pressed word delete*

[Looks in the manual for information on word delete]

Obviously, the level of detail in the user’s comments, the nature of the task and the explanations of the user are far removed from the world of games in which the child may be engrossed. In the remainder of this chapter, first the practice of giving tasks for the evaluation of computer games is discussed; should we give tasks like ‘Find the game with the rabbits and try to play it’ to the children? Subsequently, the use of thinking aloud technique is discussed. Can children really think aloud, and should they be prompted to think aloud when they forget it? Finally, in the standard usability test the facilitator only prompts the participant to think aloud, no other interaction between the facilitator and the test participant takes place. Is it possible to have no interaction between the child and the facilitator, and is it desirable? What are the consequences of giving help? These questions relate to *how* we will organise the test sessions and the answers help to define the parameters that will remain fixed in the three main experiments. These parameters will also serve as the basis for the necessary coding scheme to analyse video data presented in Chapter 3.
2.3 Does providing tasks lead to natural playing behaviour?

The use of tasks during the evaluation of work-related products is common practice and certainly has advantages. By giving tasks the evaluator tries to ensure that critical features or areas of the software are addressed. However, in the context of work-related products research has shown that the order, wording and type of tasks influence users’ behaviour (Vermeeren, 1999). For games, there are also reasons to assume that giving tasks might influence children’s playing behaviour. The major reason comes from the work of Malone and Lepper (1987). They defined the important heuristic ‘Challenge’ for good game design. According to this heuristic, providing good goals and sub goals within a game is a way to create challenge. Giving a specific task like ‘Find the game with the rabbits and try to play it’ may therefore have some risks. First of all, finding and playing the rabbit-game has become a goal for the children. Although this (sub) goal is probably also present in the game, by making it into a task it may become more important to the child to achieve this goal than achieving the other goals provided by the game. Children may therefore try to accomplish the task, even without being intrinsically motivated by the game, or they may stray from the given task because the other goals provided by the game appear more interesting. Second, it is possible that giving tasks gives away information about the game. For example, the task ‘Find the game with the rabbits and try to play it’ can give away the fact that there is a game with rabbits that can be played. The next subsection describes a study that examined the influence of tasks on the behaviour of children during the evaluation of a computer game.

2.3.1 Experimental (pilot) study: The influence of tasks

This study examined the influence of tasks on the behaviour of children playing a computer game by comparing the playing behaviour of two groups of children; one group performing tasks and one group playing freely. Four measures were defined to determine this influence: number of screens visited, number of positive and negative verbal and non-verbal indications, the number of problems found, and the number of times the children asked for help. The hypotheses for all these measures were that they could be different when children play a game with or without explicit tasks given to them by the facilitator. First, a pilot study was run to create realistic tasks for the educational computer game ‘Oscar the balloonist and the secrets of the forest’ (Transposia, 2000). The tasks were determined by observing four children playing the game without giving tasks. The types of actions that many
children performed spontaneously during these test sessions were taken as the tasks for the actual study. These tasks were:

1. The forest that you see after the introduction is how it looks in spring. Now you would like to see how the forest looks during winter. Change the season in the forest to winter.
2. Several animals live in the forest. You want to find some information about the squirrel. Try to find the explanation about the squirrel.
3. All animals in the forest leave different footprints on the ground. In order to see whether you know which animal leaves which footprint you can play a game. Find this game and play it.
4. There are several games hidden in the forest. Find the game hidden near the squirrel.

After the pilot study, a new study was set up with 17 children of eight and nine years old. These children are older than the children in the rest of this thesis because the decision about the targeted age group was not yet made at the time that this study was performed. However, by hindsight there were no reasons to assume that the possible influence of tasks would not be present for younger children. Therefore, the results of this pilot experiment were used to decide whether children should perform tasks in the remainder of this thesis.

Half of the group (8 children) performed the tasks that were derived from the pilot study, and the other half (9 children) were allowed to play the game as they liked for fifteen minutes. All sessions were video recorded.

Statistical analyses were performed to determine whether there was a significant difference between performing tasks and playing freely. Although there were no significant differences in the number of verbal and non-verbal indications, and the number of problems, the Wilcoxon signed ranks test showed that there was a significant difference between the numbers of screens children visited in the task condition and the free-play condition ($Z=-2.80, p<0.01$). In the task condition children visited around 23 screens while in the free play condition children visited only about 12 screens.

An explanation for this difference is that the children in the task condition just seemed to browse a screen to see if they could fulfil the given task with this screen. If the screen did not seem to be the right one, they quickly tried another one. Children in the free play condition did not have this task so they could take more time to look for interesting things on each screen.
A Wilcoxon signed ranks test also showed that there was a significant difference in the number of times the children asked for help in each condition ($Z=-3.67, p<0.001$). In the task condition most children asked for help between two and three times. In the free play condition none of the children ever asked for help.

An explanation for this difference is that the tasks make children more aware that they are in a test situation in which they think they have to perform well. In the free play condition they often give up playing a sub game without asking for help, whereas in the task condition they try to accomplish the given tasks by asking for help without being intrinsically motivated by the game. In this particular game, the difference between free play and tasks was very extreme. In the free play condition most children just explored the game without ever trying to play and finish a sub game, therefore they never asked for help. By contrast, the task condition did provide clear goals, which the children tried to accomplish, making them ask for help.

The major problem with giving tasks for the evaluation of computer games is that both the given task and the game can provide goals. The goals set by the tasks may interfere with the intended goals of the game and because the children feel obliged to fulfil the tasks they receive more attention than the goals provided by the game. Therefore, even though the pilot study in this experiment was explicitly meant to create realistic tasks by looking at the actions that children performed when playing freely, the tasks still influenced the children’s behaviour. Children were determined to complete the given tasks and asked for help to do this. However, when looking for fun problems a clear indication of a problem is when the child stops a sub game because he/she is no longer motivated to reach the goal provided by the sub game. By giving tasks like ‘Play sub game x’ children receive additional external motivation which may not represent the motivation provided by the game. Therefore, it is no longer possible to reliably detect real motivation problems. Although even the fact that the children are asked to play the game during the test session can be considered an external motivation, this should not be increased even further by giving explicit tasks. To avoid pressuring children to play the game care was taken to explain to the children that they were not obliged to play and that they could always stop playing the game (see section 2.1.4).

### 2.4 Verbalisation techniques with children

The verbalisation of thoughts while working with a product, commonly referred to as ‘thinking aloud’, is one of the main techniques for discovering problems in a design (Rubin,
1994, Nielsen, 1993). This means that the participants are asked to provide a running commentary of their thoughts and feelings while they are performing some given tasks with the tested product. This technique is based on Ericsson and Simon’s technique to gather verbal reports as data (Ericsson and Simon, 1984). In its cleanest form, this technique requires the test facilitator (the person conducting the test) to remain silent, whatever the participants say. Only if a participant forgets to express his/her thoughts for a certain period of time, the facilitator prompts him/her to keep on thinking-aloud. Although this technique is often cited to be used during usability tests, it is often not applied exactly as prescribed by Ericsson and Simon (2000), e.g. practitioners do not give reminders in the prescribed way and practitioners often intervene in theoretically inconsistent ways (Boren and Ramey, 2000).

A problem with trying to solicit thinking-aloud with young children is that they can have difficulty verbalizing their thoughts (Boren and Ramey, 2000). One of the reasons is that it is unnatural to talk to no-one in particular. During one of the pilot tests for this thesis a child, when asked to think aloud, responded with: ‘But to whom should I be talking, to you?’ This is a clear example of Boren and Ramey’s (2000) position that people always need a conversational partner to be able to think-aloud.

Children also forget to think aloud and need to be prompted frequently to keep talking. Unfortunately, prompting could result in children mentioning problems in order to please the experimenter, leading to non-problems being reported (Nisbett and Wilson, 1977, Donker and Reitsma, 2004). In the pilot studies there were some very clear examples of children taking several seconds trying to think of something to report to the facilitator when prompted. After this verbalisation they were silent again.

A possible solution is to instruct children to think-aloud but to refrain from prompting when they forget to talk. This may result in children feeling more comfortable and therefore reporting fewer non-problems (Donker and Reitsma, 2004). This self-initiated spoken output must then be complemented with observations of their behaviour because children sometimes forget to mention problems, or even do not realise that there is a problem and/or what the problem is. In the remainder of this thesis ‘thinking-aloud’ will refer to this self-initiated output children give during a test.

Several researchers have suggested variations of thinking-aloud to make children think-aloud without having to prompt them, for example co-discovery or constructive interaction (Nielsen, 1993, Miyake, 1986), and peer-tutoring (Hoysniemi et al., 2002). However, peer-
tutoring has not yet been evaluated experimentally in comparison to the thinking-aloud method. Co-discovery (or constructive evaluation) has been evaluated in several studies (Als et al., 2005) but the results indicate that this method does not necessarily facilitate thinking-aloud because young children do not really co-operate (van Kesteren et al., 2003) and the pairs may discuss topics unrelated to the product under evaluation (Als et al., 2005). Furthermore, the extent of the acquaintance of the children in the pairs has a large influence on the effectiveness of the method (Als et al., 2005). Because of the uncertainties about the effectiveness of these methods it was decided to focus on methods that require only one child per test session.

2.5 Giving help during evaluations

When setting up an evaluation it should be clear and consistent how the facilitator will respond to questions of the children during the evaluation. Rubin (1994) advises to assist participants only when they are very lost or confused, when they are performing a required task which makes them feel uncomfortable, when they are exceptionally frustrated and may give up, or when a bug occurs or functionality is missing. For the user tests performed in the experiments in this thesis it was decided that children could receive help from the facilitator when they asked for it repeatedly. When a child asked for help the first time, the test facilitator would only encourage the child to keep on trying. The second time a child asked for help the test facilitator would give a hint and only after the third time a child asked for help on the same topic the facilitator would explain the solution in detail. This was mainly done to establish rapport with the children. Furthermore, it is very difficult for facilitators to see children struggle with an interface, who sometimes even start to cry, without trying to comfort and help them. The assumption was that when children explicitly had to ask for help, the evaluators would get enough evidence to determine that there was a problem. After performing many evaluations with children in this manner an observation was that giving help is even more advantageous than expected, not only for children but probably even for adults. An analogy from the physical world that introduces this advantage convincingly was provided by Hartson (2005):

“Imagine this example of a problem of a wardrobe door that rubbed badly on the top of the drawer below. I adjusted the door hinges to raise the door so it didn’t rub. Success! But then the door did not stay closed properly. I guess it had been hung at a slight angle; when you close it and let go, gravity pulls the door open about a few centimetres or so. The rubbing on
the drawer below kept the door from falling open like this. The first problem (rubbing on the drawer) was obscuring the second (falling open) to the point that the second problem was not even occurring, until the first problem was solved.”

When giving help to a user (adult or child) during an evaluation, problems are bypassed temporarily. Of course these problems eventually need to be solved in the product through redesign, but giving help is a temporary solution. When a problem is solved, the user can proceed using the product, thereby being able to experience new problems that would never have been experienced if the user had failed to overcome the first problem.

Although this might result in the users solving similar problems by themselves, it is up to the evaluators to determine where parts of the game may be similar so that the same improvement should be implemented more than once. This is also important for the consistency of the game. For example, it was observed when analysing the user tests of ‘Milo and the magical stones’ (MediaMix, 2002a) that many children accidentally clicked a butterfly in the lower right corner of the screen (7 out of 25). This butterfly was meant to quit the game so the facilitator would help the children to restart the game and told them what the purpose of the butterfly was. Afterwards most children would not click the butterfly again, neither in the same screen, nor in other parts of the game. However, the evaluators should advise the developers that they should make the exit-button clear in all parts of the game in the same way, not only in those parts where the problem happened to occur.

A similar phenomenon where potential problems are obscured by other problems can also occur when experts attempt to predict problems. Some students taught to use a structured expert evaluation method, called SEEM, to predict problems in games (Baauw et al., 2005), had a hard time trying to find out the purpose of some of the sub games that they were asked to evaluate. On their evaluation forms they commented that they could not detect problems in some parts of the game because they were not able to get there. When they would have been able to ask for help, e.g. from one of the developers of the game, they probably could have identified more problems. This observation relates to the findings of Cockton and Woolrych (2001). In their study of the Heuristics Evaluation method they distinguish three types of effort required to discover problems. The easiest problems to discover are perceivable problems that can usually be discovered by looking at the display. The next easiest problems are actionable problems that can be discovered by performing a few action steps (clicks). The hardest problems to discover are constructable problems for which several interaction steps involving multiple application objects are required. The experts in their
study failed to detect 80% of the constructable problems. One of the reasons for the failure to detect these problems could be that the experts need to perform several interaction steps correctly before they can detect the problem. When the product under evaluation has usability problems that prevent the experts from performing the right steps they will not be able to detect the constructable problems anymore. Even experts may therefore benefit from being able to ask for help.

2.5.1 Providing the right help

Although giving help is a good way to detect many problems, giving the right kind of help is not trivial. Rubin (1994) and Hanna et al. (1997) advise to gradually reveal more hints to get the participants past an obstacle. However, giving the right hints is not trivial. An example from one of the test sessions performed for this thesis will illustrate this. A picture of the relevant sub game in ‘Max and the magical stones’ (MediaMix, 2002a) is given in Figure 2-1. In this game the goal is to create a dam in the brook on the right side of the screen. To build this dam a beaver throws logs into the river. The children have to push ten of these logs, to the proper place. When pushing the logs to the dam the children should not accidentally touch any of the logs that are already forming a part of the dam, with the mouse. When they touch the dam again, all the logs will flow away and the beaver will give feedback. This feedback is not always the same. Some examples of the feedback are “Breaking it down to build it up!” and “Oh no!”
Figure 2-1 The subgame where children have to build a dam by pushing logs, which the beaver will throw in the water, to the place where the dam needs to be built © MediaMix.

In the test session a boy repeatedly touched the dam, without seeming to notice that the logs were going away, or that the beaver was giving feedback. After a while the boy became frustrated and finally said to the facilitator “When is this game finished?” The facilitator tried to help him by saying: “Well, you are not supposed to touch the logs.” The boy did not understand this help at all and continued touching the logs until he got tired of the game and quit it. The reason why the boy did not understand the help became clear after reviewing the video tape of this session several times. It appeared that the boy did not really understand what the purpose of the game was. He was pushing the logs because they were thrown in the water but he did not know what he was supposed to achieve by pushing the logs. When the
facilitator provided help, she focused on the superficial appearance of the problem. She thought the boy had not understood the feedback of the logs flowing away and the beaver talking about that, and therefore tried to explain the feedback. However, the facilitator should probably have told him that he was supposed to build a dam in the right creek, and show him the exact place.

2.5.2 The adapted Interaction Cycle

The example in the previous section shows that the facilitator has to provide help on the right level to be understood by the child. To explain the different levels of help we use a version of the Interaction Cycle as defined in the User Action Framework (Andre et al., 2001) with some adaptations to make it suitable for games. This User Action Framework (UAF) is based on Norman’s theory of action model (Norman and Draper, 1986) which applies to the interaction between humans and almost any kind of product or system. It should be noted that although depicted as a cycle, it is not necessary that stages are performed in such a strict order; the different stages may sometimes overlap or be performed in parallel (Lee and Barnard, 1998). Many researchers have used Norman’s model in various ways for similar purposes (Sutcliffe et al., 2000). In the UAF, Norman’s model is used specifically as an underlying structure to help think pragmatically about usability problems. This makes it very suitable for this research. To make the Interaction Cycle of the UAF more specific for games, some of the wording is changed and one of the phases is split in order to emphasise that goals and motivation are parts of the cycle that are especially important for games. This adapted version of the Interaction Cycle was described by Barendregt and Bekker (2004) and is shown in Figure 2-2.
Figure 2-2 Adapted Interaction Cycle for games describing how the interaction between a user and a computer game takes place in terms of cognitive, motivational and physical user actions.

The first change is the wording of the phase ‘High level planning’ into ‘Determining goal’. For games goals are extremely important, especially to create challenge. In productivity applications goals are often defined externally, whereas games define their own goals. This implies also that goals should be clear and interesting to the user at all times. Some examples of games in which the goals are not always clear are ‘Oscar the balloonist and the secrets of the forest’ (Transposia, 2000) and ‘Witchstuff - With Hennie the Witch and the cat Helmer’ (Karakter Interactive, 2002). In ‘Oscar’ children can find information about different animals by clicking the animals in a forest. This is a precondition for being to be able to play some games about what the animals are eating or when they are sleeping. However, this goal is not made clear beforehand so the children do not know why they should want to collect this information.

In ‘Witchstuff’ children can click objects in the rooms of Hennie’s house. These objects then make a funny sound and show a little animation but there is no reason to click these objects other than that. One child in our test sessions therefore complained “Actually, you cannot do anything here!”
By changing the wording of this phase from ‘High-level Planning’ into ‘Determining Goal’ the necessity of providing clear and interesting goals is made clearer.

The other change in the Interaction Cycle is the addition of the phase ‘Assess motivation to go on’. While this assessment could be part of the phase ‘Evaluating Outcome’ it is important to make it more explicit. In games it is possible that although the user knows how to reach a goal he/she decides that reaching the goal is boring, too time-consuming or too stressful. According to Rouse III (2005) players do not want to have to repeat their accomplishment once they have accomplished a certain goal. If the designers have created an extremely challenging puzzle that is still difficult to complete after the players have solved it once, it should not be overused. Many curiosity problems, which will be described in section 4.5.3, relate to this phase. In productivity applications, like word processors, this phase is less likely to occur.

An example of the assessment of motivation can be found in test sessions of the educational adventure game ‘World in Numbers, The Theme Park’ (Malmberg, 2003). In this game, children have to perform three screens of arithmetic operations in order to help a ticket officer clean up his room. Some children in our tests, although able to perform the necessary arithmetic, stopped after one or two screens because they decided it took too much time to reach the end goal.

The adapted Interaction Cycle as described here can be used by facilitators to understand how they should provide help to the children during the evaluation. This will be discussed in the next section.

2.5.3 Risks of providing help in the wrong phase

When the facilitator in the example of the beaver provided help to understand the feedback, she thought there was a problem in the assessment-phase. However, the real problem was probably in the planning-phase and therefore the child did not understand the help. Another risk of providing help for the wrong phase is that potential problems may get solved unintentionally before they can get detected. For example, by saying that the child should not touch the dam again with the mouse, the facilitator could have given away the fact that the purpose of the game was to build a dam in the right brook. This depends on the exact wording and gestures of the help that the facilitator gives. When the child had been able to understand this help it would have been very difficult to discover that there was a problem with explanation of the goal of the game. So, although giving help is very useful to detect as
many problems as possible, it is very important that the facilitator is fully aware of the different phases of interaction and the impact his or her help can have on the detection of problems or on the usefulness of the provided help. When in doubt, the facilitator should make sure that earlier phases of the Interaction Cycle, for example the goal or translation are understood before giving help about the feedback. In the example of the beaver the facilitator should therefore first ask whether the child knows what he/she is supposed to do (goal). If the child answers that he/she knows that the goal is to make a dam in the right brook, the facilitator can ask whether the child knows how the dam should be built. If the child answers that he/she knows that the logs should be moved to the dam, the facilitator can ask whether the child understands the feedback of the beaver.

Of course, in the sometimes stressful situation of trying to keep a child happy and engaged during the evaluation it may be very hard for a facilitator to always keep this in mind and ensure that all questions are asked. Therefore, the evaluators also have to be aware of the Interaction Cycle while reviewing the videotapes in order to detect as many real problems as possible. In all evaluations in this project, the evaluators were acquainted with this cycle and used it to determine separate problems.

2.6 Conclusion

In this chapter first the necessary precautions to ensure appropriate ethical handling of the children in the experiments for this thesis were discussed. Subsequently, some adaptations to the traditional way to conduct usability evaluations with adults for task-based products were described in order to evaluate usability and fun in computer games with children.

The experiment in section 2.3 showed that giving tasks conflicts with the explorative nature of games and the need for intrinsically motivating goals. It was therefore decided that children should be allowed to play a game freely during all tests. Furthermore, an adaptation to the thinking-aloud technique was proposed in which children are not prompted to think-aloud, because this may make them mention non-problems in order to please the facilitator. Finally, it was discussed that the facilitator should give help to children during an evaluation in order to find as many problems as possible. The facilitator should use the phases of the Interaction Cycle to decide what kind of help the children need. When the facilitator is uncertain about the phase in which the child needs help he/she should ascertain that earlier phases are understood before giving help for later phases. These points of departure will hold for the development of the coding scheme presented in the next chapter, and for the three
Adapting the evaluation process for children and computer games

experiments in the subsequent chapters. The points of departure are not investigated in experiments in this thesis but they are kept constant. They could provide starting points for new research questions for other researchers.
3 Detecting problems in games by analysing video data

In this chapter it is described how the analysis of the video data is performed in order to detect both usability and fun problems in young children's computer games. The analysis procedure is based on a rigorous method for the analysis of video data, called DEVAN (Detailed Video Analysis). However, this method is aimed at detecting usability problems in (work-related) products for adults. The adaptations to this method in order to detect usability and fun problems with children are discussed. Furthermore, the reliability of the proposed underlying coding scheme, which is adapted for the detection of problem indicating behaviour in games, is measured.

---

2 This chapter is based on the following manuscripts (Barendregt and Bekker, 2005a, Barendregt and Bekker, 2005).
3 Detecting problems in games by analysing video data

3.1 The evaluator effect in analysing video material

When analysing video material to detect problems there is always the risk of an evaluator effect (Jacobsen et al., 1998), which means that different evaluators identify different sets of problems from the same material. Several suggestions have been made to decrease (though probably not eliminate) this evaluator effect. One suggestion is to add more evaluators (Jacobsen, 1999), another suggestion is to structure the analysis process (Vermeeren et al., 2002, Cockton and Lavery, 1999). Although a combination of these suggestions might seem an ideal solution, with increasing the complexity of the analysis process, the ratio of (session time)/(analysis time) decreases. Ratios usually lie between 1:5 and 1:100 (Sanderson and Fisher, 1994), and for example in Vermeeren et al. (2002) these ratios varied between 1:25 and 1:29. This increase in time to analyse all sessions may make it impossible to use more than one evaluator. As a basis for the analysis of the video recordings in this thesis, an analysis procedure similar to the DEtailed Video ANalysis method (DEVAN) (Vermeeren et al., 2002) was used. This method was developed to detect usability problems in task-based products for adults and should be adjusted for the detection of usability and fun problems in computer games for children. Furthermore, using the method is very time consuming.

Following Jacobsen’s advice (1999) to include more than one evaluator, it was decided that two evaluators should analyse most of the user tests. Therefore, some steps of the method need to be adjusted to make it possible to use more than one evaluator. The next sections describe the adaptations to the method.

3.2 Analysis of video data

Many structured data analysis procedures like DEVAN (Vermeeren et al., 2002) and SUPEX (Cockton and Lavery, 1999) distinguish two stages of analysis. In the first stage, observations are transcribed to an interaction table. In the second stage, the interaction is analysed in detail to locate events that indicate an occurrence of a problem.

For the analysis of most of the video data in this thesis the emphasis is on the second stage of analysing the interaction. Of the first stage only the transcription of verbal utterances was applied. An example of this transcription is given in Table 3-1.
3 Detecting problems in games by analysing video data

Table 3-1 Verbal utterances of the facilitator and the child (translated from Dutch)

<table>
<thead>
<tr>
<th>Facilitator</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>Just try to click somewhere</td>
<td>What am I supposed to do?</td>
</tr>
<tr>
<td></td>
<td>Is that the right one?</td>
</tr>
<tr>
<td></td>
<td>Just the gg (?)</td>
</tr>
<tr>
<td></td>
<td>But how can he do that?</td>
</tr>
<tr>
<td>Well, the flowers at the upper side of the screen also work</td>
<td></td>
</tr>
<tr>
<td></td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>Then let’s try those</td>
</tr>
<tr>
<td></td>
<td>Where will they come out again?</td>
</tr>
<tr>
<td></td>
<td>Yes!</td>
</tr>
<tr>
<td></td>
<td>I have to let them go through again… and then they come out over there!</td>
</tr>
<tr>
<td></td>
<td>That tastes nice, mm mm!</td>
</tr>
<tr>
<td></td>
<td>This one he never takes</td>
</tr>
<tr>
<td></td>
<td>Haaa!</td>
</tr>
<tr>
<td></td>
<td>How many are there still?</td>
</tr>
<tr>
<td></td>
<td>That’s hot when you put it there</td>
</tr>
<tr>
<td></td>
<td>I want to let them go through there</td>
</tr>
<tr>
<td></td>
<td>It’s not possible, watch!</td>
</tr>
<tr>
<td></td>
<td>Another one!</td>
</tr>
<tr>
<td></td>
<td>Jump, jump at the end!</td>
</tr>
<tr>
<td></td>
<td>That has to, that’s not possible like that</td>
</tr>
<tr>
<td></td>
<td>Why has Max found a magical stone now?</td>
</tr>
</tbody>
</table>

In the second stage the evaluators have to detect and code the behaviours that may indicate a problem. These behaviours are called ‘breakdown indications’ which relates to Winograd and Flores’ (1986) definition of a breakdown: “any conceptual discoordination between perspectives within and between people, suggesting different ways of characterizing facts and evaluating judgments.” The different types of breakdown indications are described in a coding scheme.
However, it should be noted that the codes in this coding scheme are not meant to classify breakdown indicating behaviours but only to locate them. The result of this stage of the analysis is a list of pairs of time stamps and behavioural categories, the described breakdown indications. An example of one breakdown indication is (0.00.13.36, Puzzled), meaning that at 13 seconds and 36 milliseconds the child showed or expressed puzzlement.

Because there can be multiple breakdown indications for the occurrence of one problem, this list of breakdown indications finally needs to be grouped into problems. For example, a child may say “I don’t know how to shoot the spaceship” and then may erroneously click a button to restart the game. Both indications may belong to the same problem, ‘Unclear which button is used to shoot the spaceship’. The coding of pauses which is also part of the first phase in SUPEX (Cockton and Lavery, 1999) and DEVAN in order to identify episodes was not performed explicitly because this would further lengthen the already time-consuming analysis process. However, the evaluators reviewed the tapes repeatedly to identify (streams of) episodes and used pause times implicitly to determine which breakdown indications could be grouped together.

Naturally, it is not always clear whether certain aspects of a game cause real trouble or are actually part of the fun (Pagulayan et al., 2003). For example, it can be hard to solve a puzzle in a game and the child may show behaviour that could also signal a usability problem, like for example frustration. The decision whether a breakdown indication should be considered an actual problem or part of the challenge was left to the judgement of the evaluators. Often it is quite easy for an evaluator to decide whether something is supposed to be a challenge or not. The evaluator can ask him/herself the question: “If I were a parent, would I consider this something I had to help my child with, or would I be spoiling the fun?” This is not a decision without debate; therefore it will have to be discussed with another evaluator. During the coding of the videotapes the evaluators still had to detect this behaviour without considering possible challenge. When an evaluator decided later on in the analysis that a breakdown indication should not be considered to indicate a problem the evaluator had to describe why this particular breakdown indication was not used, making sure each breakdown indication was addressed. The evaluators could then discuss their reasons for excluding a breakdown indication and if necessary decide to include the breakdown indication again by creating an additional problem report.

Thus, the evaluators had to discuss their results at two points during the analysis; after coding the breakdown indications and after clustering and describing the problems found in
Detecting problems in games by analysing video data

Each videorecording. For each video that they analysed they first had to come up with a list of breakdown indications, and later with a list of clustered problems they both agreed on. Finally, because Cockton et al. (2003) showed that using a structured problem report improves the validity (false-positive reduction) of heuristic evaluations, a similar report format as that from Lavery et al. (1997) was created. This report format was slightly modified in order to make an explicit link from each problem to the numbered breakdown indications it was based upon. A schematic overview of the analysis process with two evaluators is given in Figure 3-1. An example of how the process works is given in Figure 3-2.

![Figure 3-1 Schematic overview of the user test analysis process](image)

3-2 Example of the detection of a problem

When evaluating computer games for children both usability and fun problems can occur, and both are important to fix. However, the coding scheme of problem indicating behaviour in DEVAN is not aimed at children or computer games. A coding scheme for children’s computer games should take into account that games provide goals to create challenge, but that it is usually not necessary to reach the goals in a predefined sequence. When playing a game a child can decide that he/she first wants to explore a number of sub games without
playing them. Furthermore, fun should be an important factor besides usability; therefore some new breakdown indications to detect fun problems might be needed. Lastly it is possible that children behave differently than adults during an evaluation, making it also necessary to include new breakdown indications. Therefore, the definitions of existing breakdown indications need to be changed, new breakdown indications need to be added and some indications have to be removed.

The coding scheme and the adaptations in order to make it suitable for children and computer games are described in the next section.

### 3.3 The coding scheme

The DEVAN checklist of breakdown indication types (Vermeeren et al., 2002) is one of the most detailed checklists of breakdown indications available. However, it was not specifically created to be used for games or with children. Games differ in many ways from other products. For example, they can offer challenge as part of the fun. However, the list of breakdown indication types provides a good starting point because it is based on a cyclic task-action model inspired by Norman’s (1986) model of action. This model of action can be used to model interactions of humans with all kinds of products and can quite easily be used to model the interaction with games as well (Barendregt and Bekker, 2004). The original list is given in Table 3-2.

**Table 3-2 Definition of breakdown indication types in DEVAN (Vermeeren et al., 2002).**

<table>
<thead>
<tr>
<th>Code</th>
<th>Short description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Wrong action</td>
<td>An action does not belong in the correct sequence of actions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>An action is omitted from the sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>An action within a sequence is replaced by another action</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Actions within the sequence are performed in reversed order</td>
</tr>
<tr>
<td>DIS</td>
<td>Discontinues action</td>
<td>User points at function as if to start using it, but then does not</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User stops executing action before it is finished</td>
</tr>
</tbody>
</table>
3 Detecting problems in games by analysing video data

<table>
<thead>
<tr>
<th>Code</th>
<th>Short description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXE</td>
<td>Execution problem</td>
<td>Execution of action not done correctly or optimally</td>
</tr>
<tr>
<td>REP</td>
<td>Repeated action</td>
<td>An action is repeated with the same effect</td>
</tr>
<tr>
<td>COR</td>
<td>Corrective action</td>
<td>An action is corrected with a subsequent action</td>
</tr>
<tr>
<td>STP</td>
<td>Task stopped</td>
<td>User starts new task, before having successfully finished the current task</td>
</tr>
</tbody>
</table>

Breakdown indication types based on verbal utterances or non-verbal behaviour

<table>
<thead>
<tr>
<th>Code</th>
<th>Short description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGO</td>
<td>Wrong goal</td>
<td>The user formulates a goal that cannot be achieved with the product or that does not contribute to achieving the task goal</td>
</tr>
<tr>
<td>PUZ</td>
<td>Puzzled</td>
<td>The user indicates:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not to know how to perform the task or what function is needed for it.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not to be sure whether a specific function is needed or not.</td>
</tr>
<tr>
<td>RAN</td>
<td>Random actions</td>
<td>The user indicates that the current action(s) are chosen randomly</td>
</tr>
<tr>
<td>SEA</td>
<td>Searches for function</td>
<td>User indicates:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not being able to locate a specific function.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>To be searching for a function of which the analyst knows it does not exist.</td>
</tr>
<tr>
<td>DIF</td>
<td>Execution difficulty</td>
<td>User indicates:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Having physical problems in executing an action.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>That executing the action is difficult or uncomfortable.</td>
</tr>
<tr>
<td>DSF</td>
<td>Doubt, Surprise, Frustration</td>
<td>The user indicates:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not to be sure whether an action was executed properly.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not to understand an action's effect.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>To be surprised by an action's effect.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The effect of an action was unsatisfactory or frustrating.</td>
</tr>
</tbody>
</table>
3 Detecting problems in games by analysing video data

<table>
<thead>
<tr>
<th>Code</th>
<th>Short description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>REC</td>
<td>Recognition of error or misunderstanding</td>
<td>The user indicates: To recognise a preceding error. To understand something previously not understood.</td>
</tr>
<tr>
<td>QUIT</td>
<td>Quit task</td>
<td>User indicates to recognise that the current task was not finished successfully, but continues with a subsequent task</td>
</tr>
</tbody>
</table>

The next sections discuss the necessary adaptations in order to make the list of breakdown indications suitable for the evaluation of children’s computer games.

3.3.1 Nature of games

The list of breakdown indications of DEVAN is aimed at finding problems during user tests where the users are presented with fixed tasks, and thus many breakdown indications on this list also relate to tasks. However, it was decided not to give explicit tasks during the evaluation as was argued in Chapter 2. Therefore, the literal wording of some of the breakdown indication may make them seem not applicable for games. However, games usually do have internal goals, e.g. sub games, and these internal goals can be considered tasks although they are not given to the child by the facilitator. By replacing the term ‘task’ with ‘sub game’ the indications that refer to tasks can still be used.

When trying to code some of the gathered video material with the DEVAN coding scheme it became clear that the breakdown indications ‘Discontinues action’, ‘Repeated action’, and ‘Corrective action’ were very hard to determine. For example, is repeated clicking of a button to shoot a missile a repeated action? Or is making a character run back to a safe place a corrective action? It was decided to remove them from the list of breakdown indications. For the breakdown indication ‘Wrong action’ it was decided that the types of actions that could be considered wrong should be defined clearly. Clicking on a part of the screen that cannot be manipulated is considered to be a wrong action. Furthermore, actions that are clearly not what the child wants to do, like for example clicking a button to quit the game before the test is over, are also considered wrong actions. Later on in the analysis it may still be regarded as part of the challenge and therefore not a problem but these behaviours should be noted first.
3.3.2 Fun problems

Fun problems are a category of problems that can occur which is not covered by the DEVAN list. There are many guidelines for the design of enjoyable computer games, (Clanton, 1998, Federoff, 2002, Shelley, 2001) but there are not many theory and research based descriptions of what makes games fun or no fun. However, Malone and Lepper (1980, 1987) do provide a taxonomy for intrinsically motivating instructional environments that is based on the thorough research of what makes computer games fun. This taxonomy uses several theories of intrinsic motivation and learning (Berlyne, 1968, Csikszentmihalyi, 1975, Csikszentmihalyi, 1978, Deci, 1975, Piaget, 1951) and provides a good starting point to detect and classify fun problems (Kersten-Tsikalkina and Bekker, 2001). The taxonomy contains four main heuristics: Challenge, Fantasy, Curiosity, and Control.

- Challenge: The activity should provide a continuously optimal level of difficulty for the user.
- Fantasy: The activity may promote intrinsic motivation by using fantasy involvement.
- Curiosity: The activity should provide an optimal level of informational complexity or discrepancy from the user’s current state of knowledge and information.
- Control: The activity should promote feelings of self-determination and control on the part of the user.

Many videotapes of children playing several different computer games (Barendregt and Bekker, 2004) were used to find negative verbalisations and behaviour that could signal the occurrence of a problem when the game did not provide the right challenge, fantasy, curiosity, or control. If these indications were not yet present in the coding scheme they were added. The next subsections describe each heuristic and the breakdown indication types that children can show when the heuristic is violated. The final list of breakdown indication types will be summed up in Table 3-3.

**Challenge**

When the provided challenge in a (sub) game is too high a child will want to quit the (sub) game or ask help from the facilitator. The first indication (Quit) is already present in the original list, asking for ‘Help from the researcher’ has to be added to the list. When the provided challenge is too low the child may want to stop playing the (sub) game or become bored (Rauterberg, 1995). The first indication (Quit) is present in the list; the second
indication ‘Bored’ needs to be added. Providing the right challenge level throughout the game is also called ‘Balancing the game’ (Rouse III, 2005).

**Fantasy**

Computer games have the potential to immerse players in a fantasy world (Rouse III, 2005). However, when a child is not pleased with the provided fantasy he or she may express dislike, for example because the fantasy is too childish or too scary. The indication ‘Dislike’ needs to be added to the list.

**Curiosity**

Curiosity is aimed more at learning through playing than at playing just for fun. However, even for non-educational computer games a lack of curiosity can make a game less pleasurable to play. According to Rouse III (2005), “once players have accomplished a goal in the game, they do not want to have to repeat their accomplishment.” Children may specifically signal to be frustrated or demotivated by repetition, or a lack of progress and new experiences. Making and keeping players curious about new parts of the game is part of ‘Balancing the game’ (Rouse III, 2005). The breakdown indication ‘Bored’ added to detect possible challenge problems can also be used as an indicator for a curiosity problem. Children may also want to quit a sub game before the goal is reached because they are not curious to explore any further. The indication ‘(sub) game stopped’ can be used as a breakdown indication for this type of problem. Curiosity problems and problems related to a too low challenge level are very similar.

**Control**

Control problems may occur when the game takes over control, for example during feedback or a story, and the user cannot regain control even though he or she wants to. According to Rouse III (2005) “players expect to do, not to watch”. Although control in the sense of game mechanics is necessary to play a game, specific control problems can arise when the game takes over control for too long while this frustrates the user because he/she wants to interact. For example, a control problem occurs when the user cannot skip a long introduction or feedback, even though the user already knows the content and wants to go on playing. Children will show impatience when such a control problem occurs. The indication ‘Impatient’ should be added to the list of breakdown indications.
3.3.3 Specific behaviour of children playing games

We used preliminary versions of our coding scheme to code behaviour of children between five and seven years old, playing different adventure games, for example ‘Robbie Rabbit: Fun in the Clouds’ (MindScape, 2003), ‘Rainbow, the most beautiful fish in the ocean’ (MediaMix, 2002b) and ‘World in Numbers Group 3’ (Malmberg, 2003). While trying to code these user test data some behaviour was discovered that could not be coded with the existing breakdown indications. These breakdown indications were added to the coding scheme.

**Perception problem**

Children sometimes complained that they had difficulties hearing or seeing something properly. Examples are texts that are too small to be read or verbal explanations given in a very low volume. Another example is a situation in which the goal of a sub game is explained verbally by one of the characters in the game. Sometimes another character would talk through this explanation of the goal, making it hard to hear. The latter problem could also be called an ‘attention problem’ because the information is perceivable but the child is unable to attend to the right kind of information. However, because the coding is not performed to classify problems but just to locate breakdown indications the single breakdown indication type ‘Perception problem’ was added to the list of indications. This breakdown indication type covers all situations in which children explicitly complain about visibility or audibility.

**Passive**

Some children would stop interacting with the game when they did not know how to proceed. They would just sit and stare at the screen. Furthermore, games are often dialogues between the player and some of the characters; the child has to respond to questions and requests of the characters. However, it was not always clear to the children that an action was required of them. The children would remain passive while an action was necessary. It was therefore decided to include passivity when children were supposed to act as an indication of a problem.

**Wrong explanation**

Sometimes children at first did not seem to have a problem playing a sub game, but later they gave an explanation of something that happened which was not correct and could lead to other problems. For example, in ‘Rainbow, the most beautiful fish in the ocean’ (MediaMix,
2002b) children can decorate pictures of fishes with stamps (see Figure 3-3). When a child clicks outside the lines of the picture the chosen stamp is deactivated.

![Image of a fish decorated with stamps]

Figure 3-3 The paint studio in ‘Rainbow, most beautiful fish in the ocean’ MediaMix ©, where children can decorate pictures of fishes with coloured stamps. When a stamp is used outside the white area, the stamp is deactivated and has to be selected again.

However, one of the children in the user tests clicked outside the picture without noticing it and then said: “I’ve run out of stamps! I have to get new ones.” Because of his wrong explanation it was clear that the boy did not understand the deactivation of the stamps outside the picture. Furthermore, he kept clicking outside the picture, having to select the same stamp many times. It was decided to add giving a wrong explanation of something that happened in the game as an indication of a problem. This is similar to what Lavery and Cockton (1997) call a ‘knowledge mismatch problem’. However, it should be noted that the breakdown indication type ‘Wrong explanation’ only refers to a verbal expression made by
the child, not to a problem type as this needs further analysis. This analysis is performed after the initial detection of the breakdown indications.

3.3.4 Adaptations to eliminate overlap between indications

Some of the breakdown indications in DEVAN have two similar versions, one as an observed action on the product and one as a verbal utterance or non-verbal behaviour. When using the original coding scheme it became clear that evaluators usually code only one of the two versions. Therefore it was decided to merge these breakdown indications into one. This holds for the indications ‘Execution problem’ and ‘Execution difficulty’ and for the indications ‘Stop’ and ‘Quit’. Furthermore, the distinction between ‘Searches for function’ and ‘Puzzled’ was unclear. It was decided to remove the indication ‘Searches for function’ because the ‘Puzzled’ indication could usually cover these situations even when both occurred.

3.3.5 The proposed coding scheme

The proposed list of breakdown indication types is depicted in Table 3-3. It contains seven new breakdown indications (Passive, Impatience, Wrong explanation, Bored, Perception problem, Help, and Dislike), and six breakdown indication types are removed from the original list (Repeated action, Corrective action, Discontinues action, Execution difficulty, Searches for function, and Stop). ‘Passive’, ‘Wrong explanation’, and ‘Perception problem’ were added because children displayed this behaviour but this could not be captured by the DEVAN codes. The breakdown indication types ‘Impatience’, ‘Bored’, and ‘Dislike’ were added to capture behaviour that is an indication of a fun problem. Help was added because in the set-up of the user tests the children could ask help from the facilitator. The breakdown indication types ‘Repeated action’, ‘Corrective action’, and ‘Discontinues action’ were removed because it was too hard for evaluators to decide when to code these indications. The breakdown indication types ‘Execution difficulty’, ‘Searches for function’, and ‘(Sub) game stopped’ were collapsed with other breakdown indication types because they were actually non-verbal versions of the breakdown indication types ‘Execution problem’, ‘Puzzled’, and ‘Quit’. Evaluators often used only one of the two codes.
Table 3-3 Definition of used breakdown indication types in this thesis. The grey rows are new breakdown indication types.

<table>
<thead>
<tr>
<th>Code</th>
<th>Short description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Wrong action</td>
<td>An action does not belong in the correct sequence of actions.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>An action is omitted from the sequence.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>An action within a sequence is replaced by another action.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Actions within the sequence are performed in reversed order.</td>
</tr>
<tr>
<td>EXE</td>
<td>Execution/motor skill problem</td>
<td>The user has physical problems interacting correctly and timely with the system.</td>
</tr>
<tr>
<td>PAS</td>
<td>Passive</td>
<td>The user stops playing and does not move the mouse for more than five seconds when action is expected.</td>
</tr>
<tr>
<td>IMP</td>
<td>Impatience</td>
<td>The user shows impatience by clicking repeatedly on objects that respond slowly or the user expresses impatience verbally.</td>
</tr>
<tr>
<td>STP</td>
<td>Sub game stopped</td>
<td>The user stops the sub game before reaching the goal.</td>
</tr>
<tr>
<td>WGO</td>
<td>Wrong goal</td>
<td>The user formulates a goal that cannot be achieved in the game.</td>
</tr>
<tr>
<td>WEX</td>
<td>Wrong explanation</td>
<td>The user gives an explanation of something that has happened in the game but this explanation is not correct.</td>
</tr>
<tr>
<td>DSF</td>
<td>Doubt, Surprise, Frustration</td>
<td>The user indicates:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not to be sure whether an action was executed properly.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not to understand an action’s effect.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The effect of an action was unsatisfactory or frustrating.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Having physical problems in executing an action.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>That executing the action is difficult or uncomfortable.</td>
</tr>
</tbody>
</table>
3 Detecting problems in games by analysing video data

<table>
<thead>
<tr>
<th>Code</th>
<th>Short description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUZ</td>
<td>Puzzled</td>
<td>The user indicates:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not to know how to proceed.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not to be able to locate a specific function.</td>
</tr>
<tr>
<td>REC</td>
<td>Recognition</td>
<td>Recognition of error or misunderstanding: the user indicates to recognise a preceding error or misunderstanding.</td>
</tr>
<tr>
<td>PER</td>
<td>Perception problem</td>
<td>The user indicates not being able to hear or see something clearly.</td>
</tr>
<tr>
<td>BOR</td>
<td>Bored</td>
<td>The user verbally indicates being bored. The user non-verbally indicates being bored by sighing or yawning.</td>
</tr>
<tr>
<td>RAN</td>
<td>Random actions</td>
<td>The user indicates verbally or nonverbally to perform random actions.</td>
</tr>
<tr>
<td>HLP</td>
<td>Help</td>
<td>The user cannot proceed without help and either asks for it or the researcher has to intervene in order to prevent serious problems</td>
</tr>
<tr>
<td>DIS</td>
<td>Dislike</td>
<td>The user verbally indicates to dislike something</td>
</tr>
</tbody>
</table>

3.4 Measuring the reliability of the coding scheme

In order to determine reliability several measures are available: Cohen’s kappa and the any-two agreement measure. Cohen’s kappa (Cohen, 1960) is one of the commonly used measures that estimate the proportion of agreement between two evaluators after correcting for the proportion of chance agreement. However, Cohen’s kappa is based on each evaluator classifying the same observation points. In the case of free detection of all breakdown indicating behaviour not all evaluators may have noticed certain behaviour, resulting in different observation points (see Figure 3-4).
Detecting problems in games by analysing video data

![Table showing video analysis log file 1 and 2 comparison]

Figure 3-4 Different observation points from two different evaluators.

Furthermore, Cohen’s kappa assumes that the total number of breakdowns that need to be coded is known, or can reliably be estimated. Since it is possible that all evaluators have failed to observe certain behaviour this is probably not true. Therefore, Cohen’s kappa can not be used to determine the reliability of the scoring. Therefore, similar to Hertzum and Jacobsen (2001) who determined the reliability of problem detection, the any-two agreement measure was used to determine the reliability of the coding scheme:

\[
\frac{|P_i \cap P_j|}{|P_i \cup P_j|} \quad \text{(Over all } \frac{1}{2} n(n-1) \text{ pairs of evaluators)}
\]

(1)

In this equation, P1 and P2 are the sets of problem indications detected by evaluator 1 and 2 respectively. By replacing the sets of problems (P1 and P2) with sets of breakdown indication type/time pairs, this measure can also be used to determine the agreement of coded observation points.

The next subsections will describe three studies to determine the reliability of the coding scheme. The first study focuses on free detection of problem indications by calculating any-two agreements for four evaluators. The second study focuses on using the coding scheme to code a fixed set of observation points. Because the number of observation points is fixed it is possible to calculate Cohen’s kappa and get some insight in how well evaluators are able to distinguish the different codes. The last study investigates the benefits of training during free detection.
3.4.1 Analysis 1: Determining any-two agreements

To determine the reliability of the proposed coding scheme in terms of any-two agreement, three evaluators and one of the authors coded a piece of ten minutes videotape of a child playing a computer game called ‘Rainbow, the most beautiful fish in the ocean’ (MediaMix, 2002b). The child was asked to talk as much as possible about playing this game but was not reminded to think aloud. Furthermore, the child was not asked to perform any tasks because this is not assumed a representative way of using a game (Barendregt et al., 2003).

Before the actual coding all evaluators attended a classroom meeting in which all breakdown indication types were explained. After the explanation the individual evaluators could borrow a laptop on which the Noldus Observer™ (Noldus, 2002) was installed along with the coding scheme. They also received a CD-ROM with the game used in the user test so they could become familiar with it before the coding.

After the evaluators had completed their observations individually, all observations were compared to each other to determine the number of agreements, disagreements and unique observation points. Based on preliminary analyses with the coding scheme it was decided that observation points within four seconds of each other should be counted as the same observation points. This four-second interval was chosen because some evaluators coded behaviour after it happened while others coded behaviour by reviewing the data and going back to the beginning or the middle of a certain behaviour to code it (especially for sentences). By using four seconds these codes were still considered to relate to the same situation.

When two evaluators had the same observation point and the same code at this point, it was counted as an agreement. When one of the evaluators had an observation point and the other did not, it was counted as a unique observation for the evaluator that had coded it. When two evaluators had the same observation point but unequal codes this was counted as a disagreement. The results of each comparison are shown in Table 3-4.

Table 3-4 Number of agreements, numbers of unique observation points for each evaluator, and number of disagreements

<table>
<thead>
<tr>
<th>Eval A x Eval B</th>
<th>Any-two Agreements</th>
<th>Unique A</th>
<th>Unique B</th>
<th>Disagreements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eval 1 x Eval 2</td>
<td>50%</td>
<td>37</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>Eval 1 x Eval 3</td>
<td>33%</td>
<td>21</td>
<td>28</td>
<td>9</td>
</tr>
</tbody>
</table>
Detecting problems in games by analysing video data

<table>
<thead>
<tr>
<th>Eval A x Eval B</th>
<th>Any-two</th>
<th>Agreements</th>
<th>Unique A</th>
<th>Unique B</th>
<th>Disagreements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eval 1 x Eval 4</td>
<td>47%</td>
<td>31</td>
<td>25</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Eval 2 x Eval 3</td>
<td>27%</td>
<td>21</td>
<td>39</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Eval 2 x Eval 4</td>
<td>45%</td>
<td>31</td>
<td>24</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Eval 3 x Eval 4</td>
<td>29%</td>
<td>16</td>
<td>17</td>
<td>16</td>
<td>7</td>
</tr>
</tbody>
</table>

The average any-two agreement is 38.5%, which is rather low. However, it is in the range of any-two agreement measures reported in an overview study by Hertzum and Jacobsen (2001), although these numbers were based on problem detection instead of breakdown indication detection. A qualitative analysis was performed to determine causes for the unique observations. A major cause for unique observation points was that sometimes evaluators had not just coded all indicating behaviour but had made a decision about the severity or multiple occurrence of a breakdown. For example, when a child made the same error more than once, some evaluators had stopped coding this behaviour because they reasoned that the problem would be reported anyway. However, this was not the intention of the coding scheme because it was meant to just detect problem indicating behaviour. Reasoning about problems and multiple occurrences of problems is something that should be done in a later phase of the analysis.

It also became clear that some of the unique observation points were caused by unintended additional interpretations of the breakdown indication types. An example of an additional interpretation is that one of the evaluators had coded ‘Recognition of error or misunderstanding’ also when a child said something like: “I have been here before”, which is not a recognition of an error or misunderstanding but a recognition in general. Training the evaluators more intensively about how to apply the coding scheme could probably decrease the numbers of unique observations. This assumption is investigated more closely in section 3.4.3.

Another cause for many unique observations is that it is very hard to manually keep track of all clicks that children perform. For example, if a child clicks objects that are not clickable, the evaluator has to register each click of the child by clicking the code ‘Wrong action’ in the Observer™. When the child is fast it is very hard for the evaluators to register each click. Often one of the evaluators will miss one or more clicks resulting in unique observations. A
good solution would be if the clicks were automatically recorded and imported in the Observer™ file so that the evaluators only have to code the behaviour instead of detecting and coding it. Although missing repeated clicks had a negative influence on the any-two agreement it should be noted that these clicks would probably be grouped into one problem in a later phase of the analysis, as is also done in SUPEX (Cockton and Lavery, 1999).

Finally, it was discovered that part of the real disagreements was related to the codes ‘Impatient’ and ‘Wrong action’. When a child clicked an object rapidly and frequently it could be coded as ‘Impatient’ because it showed impatience or it could be coded as ‘Wrong action’ because it usually involved an object that could not be clicked (and therefore did not respond, resulting in impatience). Two other codes that seemed to lead to disagreement were ‘Puzzled’ and ‘Doubt Surprise Frustration’ (DSF). ‘Puzzled’ is meant for confusion before an action is executed, DSF after an action is executed. However, sometimes it is difficult to determine whether the confusion is before or after an action. For example, incomprehensible feedback can lead to confusion about the performed action but also to confusion about what is expected next. In both cases we think it is not really important which code is used, since the codes are not use for classification purposes, as long as all evaluators notice the behaviour of interest.

3.4.2 Analysis 2: Determining Cohen’s kappa for fixed observation points

To determine the extent to which unclear breakdown indication contributed to the low any-two agreement another small study was set up. In this study it was possible to calculate Cohen’s kappa properly because the evaluators received a list of observation points for which they had to pick a code. This list was created by taking the lists of all four evaluators of the first experiment. When at least three out of four evaluators in the first experiment agreed on an observation point it was included in the list of observation points that had to be coded by the new evaluators. A list of 29 fixed observation points was created in this way.

The new evaluators were two students who had received the same training as the first group of evaluators and had coded child behaviour in games with preliminary versions of this coding scheme of breakdown indications. The evaluators received the latest list of breakdown indications with explanations, a list of observation points, the game, and a CD with the video data. Independently they had to code all 29 observation points by picking one of the breakdown indications. When they thought an observation point should be coded by more
than one breakdown indication they could add these additional codes between brackets, but they had to pick one code as being the most important. Of the 29 given observation points, 26 were coded identically, resulting in a kappa of 0.87. The following guidelines are commonly used for interpreting Cohen's kappa (Robson, 1993):

- \(< 0.40 = \text{Poor agreement}\)
- \(0.40 – 0.60 = \text{Fair agreement}\)
- \(0.60 – 0.75 = \text{Good agreement}\)
- \(0.75 = \text{Excellent to perfect agreement}\)

This means that with a kappa of 0.87 the evaluators in this experiment showed excellent to perfect agreement. For the three points that were not coded identically one of the evaluators had actually given the code of the other evaluator as a secondary code. These results give a clear indication that the coding scheme does not contain unclear breakdown indicator descriptions but that evaluators use different thresholds as to when to indicate certain behaviour as a breakdown indication.

### 3.4.3 Analysis 3: The influence of training

The first two experiments showed that differences in the coding were mainly caused by unique observations of the evaluators and less by disagreements about the same behaviour. To determine whether the number of unique observations could be decreased by providing more training and discussion about the exact use and purpose of the coding scheme, we set up another exploratory study with one of the authors and a student assistant as evaluators. In this study both evaluators first coded a piece of 20 minutes videotape with the coding scheme after a short introduction of about half an hour. The results of the first coding are depicted in Table 3-5.

<table>
<thead>
<tr>
<th>Any-two Agreements</th>
<th>Disagreements</th>
<th>Unique observations 1</th>
<th>Unique observations 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>15</td>
<td>6</td>
<td>30</td>
</tr>
</tbody>
</table>

The results show that the any-two agreement in the first session was again very low.
After the first session the evaluators discussed their observations and talked about when to code specific behaviour for an hour. For example, the author explicitly explained that indicators that appear more than once for the same problem, still need to be recorded. Both evaluators now coded another 20 minutes piece of videotape. The results are depicted in Table 3-6.

Table 3-6 The number of agreements, number of disagreements and numbers of unique observation points of both evaluators for the second session.

<table>
<thead>
<tr>
<th>Any-two</th>
<th>Agreements</th>
<th>Disagreements</th>
<th>Unique observations 1</th>
<th>Unique observations 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>55%</td>
<td>54</td>
<td>5</td>
<td>21</td>
<td>18</td>
</tr>
</tbody>
</table>

In the second session the any-two agreement was doubled and it is clear that this was due to the much lower proportion of unique observations. Thus, it appears that only after one hour of additional training after a practice session the any-two agreement becomes much higher.

3.5 Discussion

Commonly used performance measures for evaluation methods like Thoroughness, Validity, and Reliability (Hartson et al., 2001) assume that the number of detected problems is countable, and this assumption will also made in the experiments reported here because this is the only way to objectively compare the results of different set-ups of a user test. However, a limitation of all experiments in which problems are counted is that problems are not very well defined and what one evaluator counts as one problem may be counted by another evaluator as two or more problems (Connell and Hammond, 1999). Determining the behaviours that indicate a problem is first step of the analysis. This step is supported by the coding scheme presented in this chapter. The coding scheme gives evaluators a way to make detection of problem indicating behaviour more reliable. Clustering behaviours into problems that do not overlap or duplicate other problems is the next step. In this second step the evaluators also have to decide which breakdown indications relate to challenge and should therefore not be treated as signalling a problem. No further tools were developed to support this second step. The evaluators in the experiments just have to discuss their results and come up with a list of problems they both agree on. However, because the same coding
scheme and discussion procedures are used for both sides of the comparisons in the experiments, the results will still be valid.

Although created with the greatest care, based on analyses of many pilot test sessions, it is not guaranteed that the coding scheme will help to detect all possible types of problems. Therefore, it is possible that some problems were missed. However, during all analyses it was never felt that another breakdown indication type, that was not present in the coding scheme yet, was needed.

The coding scheme is specifically created to code behaviour of young children with computer games in the adventure genre in order to analyse the test sessions in the experiments for this thesis. To use this coding scheme to analyse video data of other users playing games of other genres it should be checked whether all codes occur, whether additional codes are needed, and whether evaluators are able to reliably identify the relevant behaviours.

3.6 Conclusions

In this chapter the development of a coding scheme for detecting usability and fun problems in computer games for young children is described. The coding scheme is based on the DEVAN method (Vermeeren et al., 2002), and is adapted according to the theory of fun in computer games from Malone and Lepper (1987), and observations of children playing games. Six breakdown indication types were removed from the original list. Three breakdown indication types were removed because evaluators were unable to recognise these when analysing video data of children playing a game. Three other breakdown indication types were removed they were related to a very similar verbal breakdown indication type, in which case the types were joined into one. Seven breakdown indication types were added. Three breakdown indication types were added to be able to detect fun problems. Three other breakdown indication types were added because they described behaviour that children had showed when playing a game. One breakdown indication type was added because it was decided that children could ask for help and that therefore asking for help would be a clear indication for a problem. The average any-two agreement measure of 38.5% for paired comparisons of four evaluators using this coding scheme shows that thresholds for when certain user behaviour is worth coding are different for different evaluators. However, the Cohen’s kappa measure of 0.87 for a fixed list of observation points shows that the distinction between the available codes is clear to most evaluators. Furthermore, it was shown in a pilot study that training the evaluators more intensively about how to apply the coding scheme
3 Detecting problems in games by analysing video data

based on the mistakes they made in a previous coding decreases the numbers of unique observations, and therefore increases the any-two agreement considerably. The two trained evaluators from this pilot study analyzed all video data in the remainder of this thesis by means of the proposed coding scheme. Furthermore, they discussed the results of their evaluations in order to make a decision about the clustering of breakdown indications into problems and the elimination of indications related to a right amount of challenge.
4 The influence of practice on problem detection

Although user testing is often performed for first use, it may be the case that the problems detected in an evaluation with children who play a game for the first time are different from the problems that they experience after some practice. The results of an evaluation after practice may therefore merit other changes to the game than the results of an evaluation of first use. However, the change of experienced problems in a game over time has not been thoroughly investigated. In this chapter an experiment is described to investigate the change in the numbers of different types of identified problems and the severity of the problems when children practice with a game. Additionally, the changes in effectiveness, efficiency, and satisfaction are investigated. The study shows that usability problems caused by a lack of knowledge are most often present during first use. Furthermore, fun problems related to too high a challenge level may disappear after some practice, whereas fun problems caused by the game taking over control for too long occur more often after some practice. The study shows that impact severity of problems detected during first use may be higher than when children have had more practice with a game. As a result of these changes the commonly used measures of efficiency, effectiveness and satisfaction may increase with practice. Finally, the study shows that the set of most severe problems identified during first use may be radically different from the set of most severe problems identified after some practice. Because knowledge problems experienced during first use tend to be quite severe, and are often solved by asking for help from an adult, we conclude that testing first use is more important than testing after some practice. The results obtained from testing after some practice can probably be predicted and prevented by using design guidelines.

3 This chapter is based on the following manuscript submitted for publication: (Barendregt et al., Unpublished manuscript 2005b)
4.1 Practice before the evaluation

Ideally, a user test will reveal as many aspects of a product as possible that cause users trouble and that need to be modified. However, research has shown that the set of identified problems depends both on the users taking part in the user test (Nielsen, 1994), and the evaluators analysing the data (Jacobsen, 1999, Hertzum and Jacobsen, 2001). Usually, products are tested with users who use these products for the first time, but it could be expected that there are differences in users’ behaviour when they have become more familiar with a product. The first reaction of many parents and people working with children is that children learn quickly and that the problems children may encounter when they play a game for the first time will disappear spontaneously when children play the game more often. This change over time might also result in different sets of identified problems when testing first use or use after some practice. Some problems may have been overcome while other problems may arise. When performing formative evaluations in practice, not only detecting problems but also separating severe problems from insignificant problems is important because resources may not be available to correct all identified problems (Jacobsen, 1999). However, just like the numbers of certain types of detected problems may change, the overall severity estimations of problems may also change when users become more familiar with a game. Furthermore, the list of most important problems to fix might be different when testing first use or more familiar use. However, no thorough investigation has yet been performed of the exact changes in detected problems when children practice with a game. Probably there is a pattern in the kinds of problems that are likely to have been overcome and the kinds of problems that will arise when users become more familiar with a product (Prümper et al., 1992). Furthermore, computer games should be fun to play, not just the first time but preferably also after a while. Testing a game with children after they have become familiar with it, might give a different idea of the problems that need to be fixed than testing its first use. The aim of the experiment described in this chapter is to answer the question: Should we let children practice with a game before the evaluation takes place in order to find the most relevant problems that need to be fixed? To answer this question this chapter examines the differences in numbers of identified problems of different types and their severity when testing computer games for young children of adventure genre, during first use, and after they have practiced twice with the game for about half an hour. All problems will be categorised according to two frameworks that will be presented in the subsequent two sections.
4.2 Kinds of knowledge development

Although real practice effects, like learning to drive a car, or doing complex arithmetic cannot be found in such a short time the nature of many things that young children have to learn in computer games is much simpler. For example, children may not know at first that the purpose of a certain sub game is to catch all the blue flies because the explanation of the sub game is unclear. Once they have figured this out they will know what the goal is the next time they play the same sub game. This type of knowledge development belongs to the simplest types of learning in the cognitive domain according to Bloom’s taxonomy (1964) of educational activities. There are six major categories in this taxonomy, starting from the simplest behaviour to the most complex, these are:

1. Knowledge: Recall of data or information.
3. Application: Use a concept in a new situation or unprompted use of an abstraction. Applies what was learned in the classroom into novel situations in the work place.
4. Analysis: Separating material or concepts into component parts so that its organisational structure may be understood. Distinguishing between facts and inferences.
5. Synthesis: Building a structure or pattern from diverse elements. Putting parts together to form a whole, with emphasis on creating a new meaning or structure.
6. Evaluation: Making judgments about the value of ideas or materials.

The categories can be thought of as degrees of difficulties. That is, the first one must be mastered before the next one can take place. The types of knowledge development that occur when playing a game for the first time are in the categories Knowledge and Comprehension. Schneiderman (1998) discerns two types of knowledge: semantic and syntactic knowledge. Semantics are well-structured and relatively easy to learn and retain whereas syntax is arbitrary and must be memorised; e.g. the semantic notions of computer programming are much easier to learn and retain than the syntactic details of any given programming language. By replacing command languages with GUIs the amount of syntax which users
must memorise is reduced and interfaces are moved closer to the direct manipulation model which resembles how users manipulate tangible objects in the ‘real world’.

The computer games for children in this research all have graphical user interfaces and do not require much syntactical knowledge. In contrast to the more complex types of knowledge development the children will probably have gained knowledge about how to play the game after playing it only twice for half an hour.

4.3 Theory of human errors

The coding scheme described in Chapter 3 is used to detect problems when analysing video tapes. However, a single code in the coding scheme, for example ‘Doubt Surprise Frustration’, can signal different kinds of problems. Sometimes frustration can be related to not knowing what to do, but it can also be related to a too high challenge level, or to not understanding the feedback provided by the game. To understand the changes in detected problems when children play a game for the first time or when they have practiced, it is necessary to classify the detected problems. Zapf and colleagues (Frese and Zapf, 1991, Zapf et al., 1992) proposed a taxonomy of errors occurring in work with office computers combining the work of Reason (1990), Norman (1986), Rasmussen (1982), and Hacker (1986). Norman’s model of user-system interaction is created to describe interactions of humans with all sorts of systems and can also easily be applied to games (Barendregt and Bekker, 2004). Rasmussen’s classification refers to the degree of conscious control exercised by an individual over his or her activities. For games, trying to find solutions for certain challenges requires more conscious control than performing more automatic actions, such as clicking buttons, therefore Rasmussen’s model is likely to be also applicable to games. Because the taxonomy of Frese and Zapf (1991) integrates the work of Norman, Rasmussen, and Hacker it was decided to use this as a basis for analysing the differences in problem types when testing first use versus use when children have become more familiar with a game (see Figure 4-1).

On the highest level the taxonomy distinguishes four main types of problems, functionality problems, usability problems, interaction problems, and inefficiencies. Functionality problems refer to the mismatch between tasks and the program, like for example bugs. Although bugs can be present in games their detection will not change with practice of the children. Functionality problems will therefore not be investigated in this chapter. Interaction problems occur when there is a mismatch between individuals using the product, for example in organisations. Therefore, children playing a game will not face this type of
problem. Thus, for the research of how practice with computer games for young children changes the occurrence and severity of problems only usability problems and inefficiencies will be considered. Usability problems and inefficiencies, the two types of problems that do occur in a computer game for young children, have the following descriptions:

- **Usability problems**: Usability problems occur when the functionality of a program is sufficient for its execution, but there are still errors that occur. These problems can be caused by a mismatch between the user and the computer program and are explained in more detail in section 4.3.1,

- **Inefficiencies**: Behaviour is described as inefficient when the user is successful in reaching a goal that should have been reached more easily because the system does not make this more efficient way clear to the user.

### 4.3.1 Usability problems

Usability problems can be classified further according to Zapf's taxonomy for usability problems given in Figure 4.1. This taxonomy contains two dimensions. The first dimension is based on the assumption that actions are goal-oriented (Norman and Draper, 1986), and that there are three different steps in the action process to reach these goals. First, goals and plans have to be developed, then the plans have to be monitored and executed, and finally the outcomes have to be evaluated on the basis of feedback whether the goals have been reached. The second dimension comprises three levels of action regulation similar to Hacker (1986) and Rasmussen (1982). First, complex analyses of situations are regulated at the intellectual level (Rasmussen's “knowledge based level”). Second, there are actions that are regulated by action patterns that are placed at the level of flexible action patterns (Rasmussen's “rule-based level”), and third, automatic movements organised without conscious attention that are placed at the sensorimotor level (Rasmussen's “skill-based level”). Apart from the levels of action regulation, Zapf's taxonomy contains the knowledge base for regulation, which has no equivalent in Rasmussen (1982). This knowledge base for regulation has at least three aspects: knowledge of facts, knowledge of procedures, and understanding in the sense of mental models (Anderson, 1983, Gentner and Stevens, 1983). This knowledge is used to develop goals and plans.
Knowledge base for regulation

Knowledge problems

<table>
<thead>
<tr>
<th>Levels of action regulation</th>
<th>Steps in the action process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goals/planning</td>
<td>Monitoring</td>
</tr>
<tr>
<td>Intellectual level of regulation</td>
<td>Thought problems</td>
</tr>
<tr>
<td>Level of flexible action patterns</td>
<td>Habit problems</td>
</tr>
<tr>
<td>Sensorimotor level of regulation</td>
<td>Sensorimotor problems</td>
</tr>
</tbody>
</table>

Figure 4-1 A taxonomy of usability problems (Zapf et al., 1992) with two dimensions; steps in the action process and levels of action regulation.

In the knowledge base of regulation, only one type of problem is identified:

- **Knowledge problems** occur when the user is unable to carry out a certain task with the program because the design of the program is unclear about the right commands, function keys, or rules. In a computer game these problems may occur when the goal of a sub game is not explained well, or when the actions that have to be taken to reach the goal are not explained clearly.

On the intellectual level of action regulation, three types of problems are identified according to the three steps in the action process. Problems at the intellectual level are similar to mistakes as defined by Norman (1981) and Reason (1990).

- **Thought problems** occur when the product makes users develop inadequate goals and plans, or take wrong decisions in the assignment of plans and sub plans, although the user knows all the necessary features of the system. In a computer game these problems may occur when a child wants to go to a certain part of the game but because the navigation-buttons all look alike he or she unintentionally goes somewhere else.
• **Memory problems** occur when the user forgets a certain part of the plan and does not complete it, although the goals and plans were originally correctly specified. In testing a computer game for children these problems are difficult to identify because the child’s plan should be known in order to detect a memory problem.

• **Judgement problems** occur when the user cannot understand or interpret the game feedback after an action. In a computer game these problems occur for example when feedback about the progress in the game is given but the child cannot understand whether the action he or she performed is right or wrong.

On the level of flexible action patterns, three types of problems are identified according to the three steps in the action process:

• **Habit problems** occur when the user performs a correct action in a wrong situation. With adults this may happen for example, when switching from one word-processing program to another. When the user interfaces of the two products are not consistent, the user may try using the same keys to perform an action as in the old program, even though they are not working in the new program. Children playing a computer game could, for example, have habit problems when they play many games that differ in the way the navigation works.

• **Omission problems** occur when a person does not complete a well-known sub plan. In a computer game these problems may occur when the child always has to answer a question by clicking a button before being able to go to a new part of the game. When such questions are repeated often the child may eventually forget to answer them because he/she is more focused on the next step.

• **Recognition problems** occur when well-known feedback is not noticed or confused with other feedback.

Finally, at the sensorimotor level of regulation there is only one type of problem because it is empirically difficult to differentiate between planning, monitoring and feedback at this level:

• **Sensorimotor problems** are related to the required motor-skill in a product, like accidentally clicking the wrong button because it is too small or too close to other buttons.

Problems at the level of flexible action patterns and problems at the sensorimotor level are similar to action slips as defined by Norman (1981) and Reason (1990).
4.3.2 Inefficiencies

According to Zapf et al. (1992), two types of inefficiencies can be differentiated; inefficiencies because of lack of *knowledge*, and inefficiencies because of *habit*. The first type of inefficiency occurs when the user follows an inefficient path because he or she does not know a better way. The latter implies that the user uses routines although he or she knows that there are more efficient ways. Children using a game for the first time can only have knowledge inefficiencies because the game has not made clear to them that there is a better way. Habit inefficiencies could occur when children have played the game more than once.

4.4 Intrinsic motivation in computer games

Because having pleasure and fun are key factors in a computer game (Pagulayan et al., 2003), fun problems are another category of problems that can occur. Therefore, problems that undermine fun are worth examining. For this experiment, fun problems are defined as follows:

- **Fun problems**: Fun problems occur when there are aspects in the game that make the game less motivating to use, even though they are not usability problems. For example, the music can be too scary, the characters can be too childish, or the challenge level is either too low or too high.

Malone and Lepper (1980, 1987) provide a taxonomy for intrinsically motivating instructional environments that is based on the systematic research on what makes computer games fun. The taxonomy was already described in Chapter 3 and it provides a good starting point to detect and classify fun problems (Kersten-Tsikalkina and Bekker, 2001).

Based on the taxonomy of Malone and Lepper (1987) and observations of children playing several different computer games (Barendregt and Bekker, 2004), four types of fun problems were determined: Challenge problems, Fantasy problems, Curiosity problems, and Control problems. Although other, perhaps better suitable names could be given to each type of problem, the names used in this chapter will keep a strict reference to the heuristic they were based upon. In this way, the relation to the original heuristic is kept clear.
4.5 Experiment: Research questions and hypotheses

4.5.1 Usability problems

When children practice with a game they change from complete novices to more experienced players of the game. While eight different types of usability problems have been defined, Zapf et al. (1992) found that for adults using an office application there are only three significant differences between novices and experts concerning usability problems:

- Experts have significantly fewer knowledge problems than novices.
- Experts have significantly fewer thought problems than novices.
- Experts have significantly more habit problems than novices

Although children who have played a game twice for half an hour cannot be considered experts it still seems likely that they will develop their skills in the same way. Therefore, these three significant differences are the starting point for the hypotheses about the different usability problems children will face when playing a game for the first time or after they have become more familiar with the game.

Knowledge problems

Just like adults using an office application, children who play a game for the first time will probably also have many knowledge problems because they do not know how to play the game. For knowledge problems Zapf et al. (1992) found that adult users need external support in 55.9% of the cases. For children this means that they would often need help from a parent or friend to handle this kind of problems. After they have received help, it is likely that these problems disappear. Therefore children will have fewer knowledge problems after some practice when they have been able to ask help during this practice. The hypothesis is:

H1. The average number of knowledge problems per child will be higher when children play a game for the first time, than when they have had some practice with the game in the presence of somebody they can ask for help.

Thought problems

Thought problems occur when the user develops inadequate goals and plans, or when the user makes wrong decisions in the assignment of plans and sub plans, although he or she knows all the necessary features of the system. The prediction of Zapf et al. (1992) was that
experts would have fewer thought problems than novices. However, when children play a game for the first time they do not yet know all the necessary features of the game. Therefore, they cannot have many thought problems. Only when they play a part of the game twice, or use a screen more than once can they have any thought problems. Thus, it will not be possible to see a decrease in the number of thought problems when comparing the problems found during first use and experienced use. Actually, when children get to know the game they will start having thought problems, so there will probably first be an increase. Therefore the hypothesis is the opposite of the one posed by Zapf et al. (1992):

H2. The average number of thought problems per child will be lower when children play a game for the first time, than after some practice.

**Habit problems**

Although children can certainly show habit problems because they play different games, or even use other computer programs, the present study only focuses on the effects of the change in familiarity with one game. Because we will not control the change in familiarity with different programs between the two tests we cannot test any hypotheses regarding a change in the number of habit problems.

### 4.5.2 Knowledge inefficiencies

There are two different types of inefficiencies, but Zapf et al. (1992) found that for adults using an office application there is only a significant difference between novices and experts for knowledge inefficiencies:

- Experts have fewer knowledge inefficiencies than novices.

This significant difference forms the basis for the hypothesis concerning the inefficiencies children encounter when playing a game for the first time or after they have become more familiar with the game. Just like adults using an office application, children playing a game can show inefficiencies because of a lack of knowledge. For example, in many games children are unaware of map-functionalities for easier navigation. Therefore they have to navigate through many other screens to get to the part of the game that they want to go to. Often, children get annoyed because they feel it takes too much time to get to specific parts of the game, but they do not know how to navigate more efficiently. Therefore the hypothesis is:
H3. The average number of **knowledge inefficiencies** per child will be higher when children play a game for the first time, than when they have had some practice with the game in the presence of somebody they can ask for help.

4.5.3 Fun problems

Malone and Lepper’s (1987) taxonomy for the design of enjoyable computer games can be used as a starting-point for specific heuristics for computer games (Federoff, 2002). However, as far as we know, there are no clear-cut results of other research that predict what will happen to fun problems when children become more familiar with a game. Based on observations of children playing many games some expectations for this study can still formulated.

**Challenge problems**

It can be expected that there will be fewer problems relating to a too high challenge level when children have more experience with a game. Therefore the hypothesis is that:

H4. The average number of problems per child relating to **a too high challenge** level will be higher when children play a game for the first time, than when they have had some practice with the game.

Problems relating to a too low challenge level can be expected to increase in number when children have more experience with the game. Therefore the hypothesis is that:

H5. The average number of problems per child relating to **a too low challenge** level will be lower when children play a game for the first time, than when they have had some practice with the game.

**Control problems**

Control problems occur when the child cannot control the game, even though he or she wants to. This can happen especially when the user cannot skip long explanations or feedback. When children have more experience with the game they will be more impatient to take control, therefore the hypothesis is:

H6. The average number of **control problems** per child will be lower when children play a game for the first time, than when they have had some practice with the game.
Fantasy and Curiosity problems

Fantasy problems occur when the user is not pleased with the provided fantasy in a game. While positive indications of fantasy occur often when a child is playing a game, negative indications of fantasy rarely occur (Kersten-Tsikalkina and Bekker, 2001). Therefore, the chance of finding fantasy problems, which are based on negative rather than positive indications, is small. The game in the present experiment was based on a popular picture book for children (Pfister, 1999) and it was reasoned that the number of detected fantasy problems would therefore probably be too small to see any change when children would become more familiar with the game. Therefore no hypothesis is posed for fantasy problems.

Curiosity problems occur when the user has to play a sub game for too long before being able to reach a (sub) goal, get a reward, or move on to a new part of the game. On the hand it is possible that when becoming familiar with the game, children may be able to reach (sub) goals quicker after a while. On the other hand, children may also become more frustrated when they have to play long games again to visit new parts of the game. From observing children play other games it was also not clear whether children would have more or fewer curiosity problems when they practiced with a game. Therefore no hypothesis was formulated about curiosity problems.

4.5.4 Severity measures for problems

To determine whether the estimations of problem severity will change when problems are encountered during first use or experienced use, all problems encountered in both test sessions will be assigned two severity measures: Frequency Severity and Impact Severity. These severity measures will be described in section 4.9.3.

Most important problems

In the development of a computer game, design resources may not be available to correct all identified problems (Jacobsen, 1999). Therefore, a list of highest priority problems is a valuable tool to decide on efficient use of resources. Such a list could include e.g. only the problems with the highest impact severity the ten problems with the highest frequency severity. Beside checking whether overall severity estimations of a problem change when it is encountered during the first test session or the second test session, it is also worthwhile to look at changes in these lists of most important problems. Problems that occur on one list,
but not on the other, imply that the fixed product may be different when implemented changes are based on testing first use or experienced use.

4.5.5 Efficiency, effectiveness, and satisfaction

As stated before, detecting and eliminating causes for problems in a product should eventually lead to increases in efficiency, effectiveness, and satisfaction. As hypothesised, children will be able to overcome some problems, like knowledge problems and high challenge problems when they can practice with a game. Therefore, the efficiency, effectiveness, and satisfaction measures may also change when children become more familiar with a game. Most games for young children consist of a number of sub games that have to be finished in order to achieve the highest goal, e.g. helping an animal friend to find his way home, rescuing a princess from an evil wizard etc. In this study Effectiveness is measured in terms of the number of successfully finished sub games within the duration of the test session. Efficiency is measured as the number of successfully finished games divided by the number of visited games (this means that children really play and finish the sub games that they visit and do not visit many games that they cannot play). Satisfaction is measured by asking children how much they like the game with a Likert-scale instrument developed to measure fun with young children (Read et al., 2002). More information about this instrument is given in section 4.6.3.

4.5.6 Previous experience

Previous experience with computer games may influence the changes in problem types when children become more familiar with a game. For example, it is possible that experienced children have fewer knowledge problems when they play a game for the first time than inexperienced children. When they get more familiar with the game they may not show the same decrease in knowledge problems as less experienced children. To determine whether previous experience with computer games is of influence on the changes in numbers of identified problems of each type, an explorative analysis will be performed. Experience is measured as the average number of times a child uses the computer per month, and the average duration each time the child uses the computer. These measures are combined to an overall experience scores which is used as the independent variable in the regression analyses for the number of detected problems of each type in the first session. Second, the experience score will be used as the independent variable in the regression analyses for the number of
detected problems of each type in the last session. Finally, the experience score will be used as the independent variable in the regression analyses for the difference in the number of detected problems of each type in the first and the last session.

4.6 Method

4.6.1 Participants

To test the hypotheses an experiment was initially set up with 26 children of group three and four (grade one and two) of De Brembocht, an elementary school in Veldhoven, The Netherlands. This school is situated in a neighbourhood that is mainly inhabited by people who received higher education and earn more than minimum wage. All children were between five and seven years old (mean age was 84 months, SD=5.6 months), 9 girls and 17 boys. Unfortunately, one of the 26 children was ill and could not perform the second session. Therefore there were only 25 children in the experiment.

The children were recruited by a letter to their parents asking for their cooperation. Prior to the test the parents had to complete a small questionnaire concerning the frequency with which their child used the computer and the average time their child used the computer. The combined results of this questionnaire are given in Figure 4-2.
Figure 4.2 Numbers of children who use the computer almost never, once per month, once per week, several times per week, or (almost) every day, subdivided by the time that they use the computer.

4.6.2 Test material

The 25 children in the experiment took part in two identical facilitated test sessions (see 4.6.3 for an explanation of the test sessions) in which the children played ‘Milo and the Magical Stones’ (MediaMix, 2002a). The game could not be saved in the first session so the children had to start again from the beginning in the second session. Between the two facilitated sessions the teacher made sure that each child had an additional half hour of free play with this game in the classroom.
4.6.3 Procedure

The nature of the types of game tested in this thesis is so that a child who has played a game for more than an hour can already be classified as rather familiar with the game (see section 1.6). In this study it was therefore decided to have the children play the game three times, of which the first and the third times that they played the game were examined. In the first test session the children played the game for the first time for half an hour. In the following three weeks they were allowed to play the game for a second time for half an hour. Therefore, the third time the children played the game they had practiced with the game for about one hour in total (half an hour twice). The children were not allowed to save the game, so when they played the game in the second and the third session they had to start anew, making sure they were familiar with at least some parts of the game they would be playing during the last session.

Each child was taken from the classroom separately for 30 minutes to perform the two test sessions with the game. At the beginning of the test session, the test facilitator explained the purpose of the test and instructed the child to try to talk-aloud. The child could play the game as he or she liked without any specific tasks, the test facilitator did not remind the child to talk-aloud during the test, and help was given as described in Chapter 2. After 20 minutes the facilitator would tell the child the test session was over but that he or she could choose to continue playing the game for another five minutes or return to the class. If the child chose to continue playing, the session was stopped after these extra five minutes. In the next three weeks the child played the game once more without the presence of a facilitator. After three weeks, the last test session was performed with all children. The procedure for this last test session was precisely the same as for the first session. So again, the child could play the game as he or she liked without any specific tasks, the test facilitator did not remind the child to talk-aloud during the test, and help was given as described in Chapter 2. After 20 minutes the facilitator would tell the child the last test session was over but that he or she could choose to continue playing the game for another five minutes or return to the class. If the child chose to continue playing, the session was stopped after these extra five minutes.

Both the first and the last session were videotaped, recording a split-screen shot of the face of the child and the on-screen actions.

To determine whether the satisfaction of the children changed when they became more familiar with the game they were also asked to rate the game after both the first and the last
session on a five-point scale smileyometer that was developed by (Read et al., 2002) to measure fun with children aged between 5 and 10 (see Figure 4-3). Although young children tend to pick the 'best' face, making it not a very good tool to discriminate between different programs (Read, 2005), it was used here to determine whether there was a change in opinion about the same game after the two sessions. The children were asked to answer the question ‘How much did you like the game Max and the magical stones?’ by ticking one of the boxes under the faces of the smileyometer at the end of each test session.

![Smileyometer Faces](image)

Figure 4-3 Dutch version of the smileyometer (Read et al., 2002) used in this experiment.

### 4.7 Analysis of the test sessions

#### 4.7.1 Introduction

For detecting and coding the video data in order to detect problems, Observer Pro (Noldus, 2002) was used, a software package for observational research. With this software observations can be logged with the digital video data by clicking the appropriate behavioural category as defined in the coding scheme (see section 3.3). The result of this stage of analysis is a list of time stamps combined with a behavioural category, the breakdown indications described in section 3.3.5. An example of one breakdown indication is \(0.00.13.36, \text{Puzzled}\), meaning that at 13 seconds and 36 milliseconds the child showed or expressed puzzlement.

Because of concern for the evaluator effect, eight out of the 25 user tests obtained for the first session were analysed by two evaluators. Furthermore, the two evaluators discussed all problems extensively.

For the analysis of the video data of the second test session it was decided to use a procedure with two evaluators in which they did not have to log everything as extensively as in the analysis of the first session, this procedure is described in section 4.7.2.
4.7.2 The analysis procedure for the last test session

Two evaluators analysed all videotapes of the second session according to an analysis procedure that required less extensive logging than for the first session. By making the analysis procedure for the longer-term use session less time-consuming than for the first test session it was possible to have all videos analysed by two evaluators, instead of only a part. Performing and discussing the results of all test videotapes with two evaluators was considered more important than logging everything. Furthermore, it was reasoned that the evaluators would have enough experience after analysing the sessions for first use to recognise relevant behaviour that signals a problem in the sessions for experienced use.

Instead of explicitly scoring breakdown indications with the Observer, the evaluators started with the total list of problems found in the first session and marked the occurrence of existing problems for each child by looking at the videotapes. To identify these problems they still made use of the breakdown indication types checklist described in section 3.3.5. When they encountered a problem that did not yet occur on the list, they created a new problem description. After analysing the videotapes the evaluators discussed their results to arrive at a single agreed-upon list of problems for each child just like they had done for the first session.

4.8 Reliability analyses

4.8.1 Reliability of the problem analysis of the two sessions

To check the inter-coder reliability for the first session any-two agreement measures were calculated for the results of the individual breakdown coding. The average any-two agreement measure for the eight analysed user tests of the first session was 0.73. The any-two agreements ranged from 0.49 (for a user test with a boy who clicked very frequently making it difficult to keep up with all the actions), to 0.90 (for a user test with a boy who performed the same actions over and over again making it rather predictable what would happen). The average any-two agreement measure for detected problems in all 25 analysed videotapes of the second session was 0.68. Both values are about the same as Vermeeren et al. (2003) found in several user tests of different types of products.

4.8.2 Reliability of the problem classification

Two evaluators independently classified all problems of the first and last session according to the problem taxonomies described in section 4.3 and 4.4. Of all 146 problems they classified
136 problems in the same high-level category (usability problem, fun problem, or inefficiency), resulting in a Cohen’s kappa of 0.81. Of the 113 problems that were classified as usability problems, 103 were classified at the same lower level (knowledge problem, thought problem, judgement problem, omission problem, or sensorimotor problem) resulting in a Cohen’s kappa of 0.79. Of the 19 problems that were classified as fun problems, 16 were classified at the same lower level (challenge problem, control problem, fantasy problem, or curiosity problem), resulting in a Cohen’s kappa of 0.81.

This means the observers in the experiment showed excellent to perfect agreement for both the high-level classification and the low-level classification. Afterwards the evaluators discussed all problems that were classified in different categories and created a final classification of all problems that they both agreed on.

4.9 Results

4.9.1 Types of problems in both test sessions

In the first test session 98 problems were identified, and in the second test session 115 problems. Some of the problems identified in the first session were the same as problems identified in the second session. Others were only identified in one of the two sessions. For each session the distribution of the identified problems in this session over the problem categories is given in Figure 4-4.
4.9.2 Hypotheses problem types

Most children did not visit exactly the same parts of the game in the first and the last test session. In the first session they visited on average seven different screens of the game, while in the last session they visited on average nine different screens of the game. On average six screens were visited in both sessions. For those parts of the game that a child visited both times each child identified on average 15 problems in the first session and only 8 problems in the last session. A paired $t$-test shows that this difference is significant ($df=24, p<0.0001$). When children played a specific part of the game more than once they experienced fewer problems in this part of the game. The average numbers of identified problems of each
problem type per child for those parts of the game that the children visited in both sessions are given in Table 4-1.

Table 4-1 Analysis of variance (paired t-tests) for the number of problems of each type per child for first use and after an hour of practice.

<table>
<thead>
<tr>
<th>Problem type</th>
<th>Mean nr. of problems per child in first session (n=25)</th>
<th>Mean nr. of problems per child in last session (n=25)</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Inefficiencies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge inefficiencies</td>
<td>0.88</td>
<td>0.92</td>
<td>-0.161</td>
</tr>
<tr>
<td>Habit inefficiencies</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
</tr>
<tr>
<td>Usability problems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge problems</td>
<td>9.48</td>
<td>4.04</td>
<td>9.106***</td>
</tr>
<tr>
<td>Thought problems</td>
<td>0.32</td>
<td>0.64</td>
<td>-1.398</td>
</tr>
<tr>
<td>Memory problems</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
</tr>
<tr>
<td>Judgement problems</td>
<td>1.40</td>
<td>0.68</td>
<td>2.979**</td>
</tr>
<tr>
<td>Habit problems</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
</tr>
<tr>
<td>Omission problems</td>
<td>0.00</td>
<td>0.08</td>
<td>-</td>
</tr>
<tr>
<td>Recognition problems</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
</tr>
<tr>
<td>Sensorimotor problems</td>
<td>0.00</td>
<td>0.04</td>
<td>-</td>
</tr>
<tr>
<td>Fun problems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fantasy</td>
<td>0.00</td>
<td>0.08</td>
<td>-</td>
</tr>
<tr>
<td>Curiosity problems</td>
<td>0.08</td>
<td>0.12</td>
<td>-0.70</td>
</tr>
<tr>
<td>Control problems</td>
<td>0.64</td>
<td>1.56</td>
<td>-2.963**</td>
</tr>
<tr>
<td>Challenge (high) problems</td>
<td>0.56</td>
<td>0.20</td>
<td>2.377*</td>
</tr>
<tr>
<td>Challenge (low) problems</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
</tr>
</tbody>
</table>

* Because there was no variance in one or both of the sessions, a paired t-test could not be computed.

* p<0.05, one-tailed. ** p<0.01, one-tailed, *** p<0.001, one-tailed.

H1. A one-tailed paired t-test confirmed the hypothesis that on average the number of knowledge problems per child in the first session (mean=9.48, SD=3.03) would be significantly higher (df=24, p<0.001) than the number of knowledge problems per child in the last session (mean=4.04, SD=2.79).
H2. A one-tailed paired $t$-test did not confirm the hypothesis that on average the number of thought problems per child in the first session would be significantly lower than the number of thought problems per child in the last session. However, the change in the average number of thought problems (0.32 in the first session vs. 0.64 in the second session) was in the expected direction.

H3. A one-tailed paired $t$-test failed to show ($p>0.05$) that the number of knowledge inefficiencies per child in the first session (mean=0.88, SD=1.01) was significantly higher than the number of knowledge inefficiencies per child in the last session (mean=0.92, SD=0.95).

H4. A one-tailed paired $t$-test confirmed the hypothesis that the number of problems related to too high challenge per child in the first session (mean=0.56, SD=0.65) was significantly higher ($df=24, p<0.05$) than the number of high challenge problems per child in the last session (mean=0.20, SD=0.5).

H5. No paired samples $t$-test could be computed for the number of problems related to too low challenge because there were no problems of this type in both sessions were observed. Therefore, no conclusion can be drawn for this type of problem.

H6. The one-tailed paired $t$-test confirmed the hypothesis that on average the number of control problems per child in the first session (mean=0.64, SD=0.64) would be significantly lower ($df=24, p<0.01$) than the number of control problems per child in the second session (mean=1.56, SD=1.53).

Although some of these differences may appear quite small, e.g. between the number of high challenge problems in the first and the last session, it should be noted that these are averages per child. This may mean that in the first session about half of all children had one problem of this type (mean=0.56), whereas in the last session only twenty percent of the children had one problem of this type (mean=0.20) in the screens that they had visited before.

An unanticipated result was that the average number of judgement problems per child was also significantly higher ($df=24, p<0.01$) in the first session (mean=1.4, SD=1.3) than in the last session (mean=0.68, SD=0.85). This is probably due to the fact that children had fewer knowledge problems. Typically, knowledge problems occur when children have a problem determining what is expected from them. In that situation they are often also unable to correctly interpret the feedback they get, resulting in a judgement problem. When children know what is expected from them, they are also able to interpret the offered feedback, resulting in fewer judgement problems.
4.9.3 Problem severity

**Severity of problems identified in both sessions**

Of the 98 problems identified in the first session 68 problems appeared again in the last session. In order to determine whether there would be a difference if the severity of the problems was determined based on the findings in the first session or the last session, two severity measures were determined: Frequency severity and Impact severity. Frequency Severity relates to the percentage of users that would encounter a problem, while Impact Severity relates to the consequences a problem will have for the user. These measures are similar to many commonly used measures to determine criticality of problems (Rubin, 1994, Nielsen, 1993). In practice both measures will often be used in combination (Rubin, 1994) to determine the overall severity of each problem. However, they will be discussed separately in this chapter to provide a clear overview of the situation. For frequency severity the number of children experiencing the problem is used, for impact severity an average of the impact the problem had on each child experiencing the problem is used, which will explained in the next subsection.

**Impact Severity**

To take into account the fact that the impact of a problem can differ from person to person, the Overall Impact Severity for each problem was determined by averaging the impact of a problem over all children who experienced this problem:

- For each child that experienced the problem a score of 1 was given if the child could continue without further help.
- A score of 2 was given if the child could continue with a little help.
- A score of 3 was given when the problem made the child deliberately quit a subgame with or without being given help, or when the facilitator had to take over.

The severity of each problem was calculated by adding the scores for all children, and dividing this total score by the number of children who experienced the problem. The highest possible score would therefore be 3.0, the lowest possible score 1.0. For example, suppose a certain problem was experienced by two children. The first child could overcome the problem without any help; the second child could overcome the problem with some help from the facilitator. For the first child the impact severity would be 1, for the second child the impact severity would be 2. The Overall Impact Severity of this problem would thus be 1.5.
4. The influence of practice on problem detection

**Severity of the detected problems**

For each problem the two severity measures were determined twice, once for the occurrence of the problem in the first test session, and once for the occurrence of the problem in the second test session. Figure 4-5 shows a scatter plot of the *frequency severity* for each problem in both sessions.

Figure 4-5 Frequency severity for problems in both the first and the second session (n=68). The possible frequency severity ranges from 1 to 25. The numbers at each data point show the number of problems. Data points below the dotted line mark problems that decrease in frequency severity from the first to the second session. Data points above the dotted line mark problems that increase in frequency severity.

Figure 4-6 shows a scatter plot of the *impact severity* for each problem in both sessions.
Figure 4-6 Impact severity for problems in both the first and the second session (n=68). The possible impact severity ranges from 1.0 to 3.0. The numbers at each data point show the number of problems. Data points below the dotted line mark problems that decrease in impact severity from the first to the second session. Data points above the dotted line mark problems that increase in impact severity.

A paired t-test showed that the frequency severity of problems was not significantly different ($p>0.05$) between the first and the last session. The impact severity of problems was significantly lower ($t(67)=2.148$, $p<0.05$, 2-tailed) in the last session (mean=1.23, SD=0.51) than in the first session (mean=1.35, SD=0.56) according to a paired t-test. This corresponds with the findings of Prümper et al. (1992) who reported that experts spend less time handling errors than novices. Error handling time can be viewed as an alternative measure for impact severity because both measures express the negative effects of a problem for the user.
Most important problems

The result of each user test is a large set of problems discovered in the game. However, for the game designer, programming resources may often not be available to correct all identified problems (Jacobsen, 1999). Therefore, the developers have to decide which problems they will fix first and which they will fix later, or leave unattended. This decision is usually based on the severity estimations of the detected problems; the most severe problems are fixed first. However, just like the types of detected problems change when children become more familiar with a game, the most severe problems detected during first use are not necessarily the same as the most severe problems detected during more experienced use.

To illustrate this effect, imagine that the developers of the game used in this study would decide to fix only the ten problems that score the highest on frequency severity, and the ten problems that score highest on impact severity. For both types of severity, two sets of the ten most severe problems can be made; one set for the first session and one set for the second session.

Of the ten highest frequency severity problems from the first session, only four are also present in the ten highest frequency severity problems from the second session. Of the ten highest impact severity problems from the first session, again only four are also present in the ten highest impact severity problems from the second session.

Thus, six out of ten problems with the highest priority according to user test of first use are no longer the most important problems when testing experienced use, and vice versa. This means that it would make a huge difference when the developers would change the game based on the results of the first or the second user test.

There are several reasons for this difference in importance of problems. One reason is that children may encounter problems in the first session that are so severe that they cannot proceed in the chosen direction. After they have practiced a little, and asked help from others, they may be able to proceed and experience other severe problems. The problems that were the most severe in the first session are thus replaced by new severe problems.

Another reason is that many knowledge problems have a high severity when they are encountered for the first time. Often, children have to ask for help when they encounter these severe problems. When the child receives help, the problems cease to exist, while other problems that were less severe in the first session remain almost equally severe because children do not always ask help for less severe problems.
4.9.4 Effectiveness, efficiency, and satisfaction

To determine whether the effectiveness changed when children became more familiar with the game the number of finished sub games within the half hour of the test session per child was compared for the first and the last test session. A Wilcoxon signed ranks test showed that the average number of finished sub games per child in the last session was significantly higher ($Z=-3.96, p<0.001$) than the average number of finished sub games per child in the first session. In the first session children on average were able to finish two games, in the second session they were able to finish four games.

Although children experienced more control problems in the last session, this type of problems does not take any time to overcome. Control problems are merely annoyances because children want to proceed faster than the game allows them to. Therefore these problems did not have an effect on the effectiveness but they could have had a negative effect on the satisfaction.

To determine whether efficiency changed when children became more familiar with the game the number of finished sub games per child divided by the number of visited sub games was compared for the first and the last test session. The two-tailed paired $t$-test showed the average proportion of finished sub games per child in the last session (mean=0.31, SD=0.19) was significantly higher ($t(24)=-2.910, p<0.01$) than the average proportion of finished sub games per child in the first session (mean=0.44, SD=0.18).

A Wilcoxon signed ranks test performed on the scores that the children gave on the smileyometer after each session about how much they liked the game (1=awful to 5=brilliant) showed that there was also a significant difference in the children’s satisfaction in playing the game after the first and last session ($Z=-2.486$ based on negative ranks, $p<0.01$). The children appreciated the game more after the last session. This trend is in agreement with the fact that fewer knowledge problems and problems related to too high challenge were found in the last session than in the first session.

4.9.5 Experience

For each child the frequency (on a 5-point Likert-scale) and time they used computers (on a 5-point Likert-scale) as indicated by the parents, were used as input to calculate the average number of minutes of computer use per month for each child. First, regression analyses were performed for this experience score on the number of detected problems of each type in the first session. Second, regression analyses were performed for the experience score on the
The number of detected problems of each type in the last session. Finally, regression analyses were performed for the experience score on the difference in problems of each type in the first and the last session.

For the first session there was no significant effect for experience on the number of detected problems of any type. For the last session there was only a significant effect for experience on the number of detected control problems ($t(24)=2.340, p<0.05$). This means that more experienced children had more control problems in the last session than less experienced children. Finally, for the difference between the first and the last session there was also only a significant effect for experience on the difference in the number of detected control problems ($t(24)=-2.142, p<0.05$). This last finding is of course a direct result of the combination of the two previous findings. So, when children were more experienced with computers they were also more impatient with the game when they became more familiar with it, resulting in more control problems.

### 4.10 Discussion

#### 4.10.1 Considerations about the unconfirmed hypotheses

Contrary to the expectations, the number of knowledge inefficiencies was not significantly higher for the first test session than for the last test session. This was caused in both test sessions by the low number of inefficiencies that all were identified by only a few children. The set up of the problem detection in this study makes it hard to detect inefficiencies. Only when children expressed frustration or asked help for an inefficiency could the evaluator note this as a breakdown indication. For example, in the last test session a boy played a game that he had already played in the first session. In this game the player had to make the mouse Milo catch flies while jumping from one leaf to another. The leaves were positioned in a row across a lake. In the last session the child asked: “Is it also possible to catch flies from somewhere in the middle?” The evaluators could now note this as a breakdown indication. Although the boy probably applied this tactic in both sessions it was only identifiable in the last session when the boy talked about it. Therefore, this inefficiency was not reported in the analysis of his first session.

The hypothesis on thought problems was not confirmed for only those parts of the game that the children visited in both analyzed sessions. Although the average number of thought problems was higher for the second session than for the first session, as was expected, this difference was not significant.
For challenge problems only the hypothesis on high challenge problems was confirmed. The hypothesis for the number of low challenge problems could not be tested because problems of this type were not identified in either test session. This corresponds with the findings of Bekker and Kersten-Tsikalkina (2001), who reported that children did not show any evidence of too low challenge levels with a game called ‘Junior Detectives’. Most games actually tend to be too difficult instead of too easy. In his book about game design Richard Rouse III (2005) wrote: “While balancing your game you should keep one rule of thumb in mind at all times: your game is too hard.” Although it is theoretically possible that a game is too easy, it was still more likely that the children playing the game in this experiment would find it too hard. Therefore, it is not surprising that they did not experience any problems relating to a too low challenge level, even after they had practiced with the game.

4.10.2 Limitations

The results of this experiment are based on the same group of children playing the game twice in the presence of the researchers. It is possible that the children behaved differently in the last session, not only because they were more familiar with the game, but also because they were more familiar with the researchers and the test situation. By performing the second session with a new facilitator it could have been possible to exclude this effect. However, the change in facilitator would also have introduced an extra confounding variable.

Many of the usability hypotheses were based on the results of Zapf and associates who performed their research with adults using office applications in field settings. The fact that some of their results also hold for children who have played a game for just one hour is encouraging for the generalisability of both the experiment of Zapf and colleagues and the experiment presented in this chapter.

There are two main differences between the work of Zapf et al. (1992) and the current study. First, in the present study the focus is on the change in types of problems identified during first use and after children have practiced about one hour with the game. Zapf focused on the difference between novices, who have worked with a computer system less than six months, and experts, who had worked with a system for more than three years. The difference in experience between the first and last test session in the present study was much smaller than the difference in experience between novices and an experts as studied by Zapf et al. (1992).

When children become real experts with a game, they may also experience new problems on
lower levels of the usability problem taxonomy (flexible action patterns and sensorimotor level), whereas these were almost absent in the present study.

Second, the present study focused on the implications for the types of problems identified in a test, whereas the work of Zapf focuses on an overall change of experienced problems. The limited time available for a test aimed at finding problems in a product implies that only a limited part of the product can be tested. When children play a game for the second or third time in a test session, they may visit different parts of the game and commit new errors on the highest levels of the problem taxonomy.

A limitation of this research is that the severity categorisations are specifically chosen for the problems that were detected in the specific set up of the experiments. Thus, it is based on the assumptions described in Chapter 2 that the children could ask for help and did not have to perform any tasks. Asking and receiving help and quitting a sub game were used as aids to determine the Impact Severity of a problem. The fact that children could play the game as they liked resulted in having no problems that were found by more than 65% of the children. This led to the Frequency Severity classification used in this thesis. Without these assumptions a different classification of the severity of problems would have been needed.

4.10.3 What else can be learned from a comparison of both test sessions?

By comparing the numbers of children who visit different parts of the game during both first use and after practice, it is possible to get an impression of the popularity of the different parts. Users learning to use a system through interaction, for example a computer game, will form a working model of the system (Norman and Draper, 1986). During first use of a computer game they will start to build up a mental model of this game, although this model is unstable and incomplete (Norman, 1983). During first use, children are likely to just explore parts of the game that they can reach without knowing whether these parts will be fun or not, or whether these parts are an essential piece of the puzzle or not. After first use children will have built up a certain mental model of the game, and they will also know which parts of the game are fun or essential. When children are more familiar with a game they are likely to employ this knowledge to visit only those parts of the game that they like. For example, in ‘Milo and the Magical Stones’ there were only two parts of the game that were visited by a smaller number of children in the second session. The first part was a screen where children were told they could listen to stories. These stories were to reveal some clues about the game.
but the children did not understand this purpose and they thought the stories were not entertaining. The second part was a paint studio where children could decorate their own stones. Decorating stones was not an essential part of the game to reach any goals. In the second session only a few children visited the paint studio again. To assess how popular parts of the game were, the numbers of children that visited parts of the game in each session could be used.

These examples show that popularity of sub games or other parts of the game can be determined reasonably objectively by comparing the numbers of children that visit them during first use and after they have become familiar with a game. However, it takes much effort to test a sufficient number of children in both conditions to obtain these effects. As Pagulayan et al. (2003) wrote, subjective assessments like popularity can probably be measured more easily by conducting inquiry techniques.

**4.11 Conclusions**

The experiment described in this chapter examined the differences in the outcomes of a test when a game is tested when children play it for the first time or when they have become more familiar with the game. The main research question for this experiment was: Should we let children practice with a game before the evaluation takes place in order to find the relevant problems that need to be fixed?

In the first test session, when children had no experience with the game, many knowledge problems were identified. To overcome knowledge problems users often need external help (Zapf et al., 1992); for children this means they will often need the help from a parent or sibling. In the experiment the children often received help from the facilitator for this type of problem. After they receive help the children are able to continue playing the game. However, because the aim of the developers of a game is to make their game playable for children without having to ask help from their parents, it is very important to identify this type of problems in a test. Testing first use helps to identify these problems quite well, while some of these problems may no longer be observed when testing after some practice. Furthermore, the impact severity of these problems may be underestimated when only looking at children who have become familiar with the game. In order to detect knowledge problems that need to be fixed, the evaluation must involve children who have no experience with the game. Evaluating the game with children who have played the game before, thereby having been enabled to ask help from others, may miss or underestimate these problems.
Although the number of control problems increased after some practice, especially for more experienced computer users, it is may not be necessary to evaluate the game with children after they have practiced. There are two reasons for this conclusion. First, the increase in the occurrence of this type of problem was not annoying enough to lower the overall satisfaction with the game. Second, this type of problem may be simple to predict based on some more observations of children playing games. These additional observations could be translated into guidelines that can help designers to prevent this type of problem.

In the remainder of this thesis we will focus on problems detected during first use of a computer game.
In this chapter an experiment is described to determine which personality characteristics can be used to predict whether a child will make an effective participant in a user test, both in terms of the number of identified problems and the percentage of verbalised problems. Participant selection based on this knowledge can make user testing with young children more effective. The study shows that the personality characteristic Curiosity influences the number of identified problems; a combination of the personality characteristics Friendliness and Extraversion influences the percentage of verbalised problems. Furthermore, the study shows that selection of children based on these criteria still leads to finding a representative sample of the product’s problems.

This chapter is based on the following manuscript accepted for publication: (Barendregt et al., 2005a)
5 Personality characteristics as predictors of participant effectiveness

5.1 Differences between children

There are large differences between children (and probably also between adults) in the amount of self-initiated spoken output they generate; some children spontaneously give numerous comments about problems while others keep silent throughout the whole session. This happens even when they have comparable experience playing computer games and when the test facilitator behaves the same towards all children. Recently, Donker and Reitsma (2004) found that only 28 out of 70 children made any remarks during testing under similar circumstances, and Donker [personal communication] also observed that some of the children were much more talkative than others. Furthermore, just like adults (Virzi, 1992), some children help to reveal a lot of problems, either verbally or non-verbally, while others help to reveal only a few.

If it could be possible to predict more precisely which children will make effective participants, one could do much more cost-effective high-quality testing by selecting these children. Furthermore, when an observable problem is accompanied by verbalisations or other explicit indications of a problem, the number of breakdown indications per problem increases, making it more likely that a problem will be detected by multiple evaluators (Vermeeren et al., 2002). By selecting children who will verbalise the problems that they encounter the reliability of the analysis can be increased.

It is certainly not uncommon to think that differences in personality characteristics may be effective indicators of how well a child will be able to participate in a user test. For example, Donker and Markopoulos (2001) reasoned that extraversion might significantly increase the likelihood that children voice their thoughts about usability problems they encounter, and therefore tend to increase the number of found problems. Unfortunately, they were not able to confirm their hypothesis. Another example is the evaluation of game concepts by Hanna et al. (2004). They selected children who were characterised by their parents as not shy. However, Hanna et al. never validated the assumption that these children would be better participants. Furthermore, no attention was given to the fact that it is essential that the selected children do not find radically fewer or different problems than the children who make less good participants.

In this chapter an experiment is described to determine whether certain personality characteristics of children are predictive for both the number of problems and the number of spontaneous comments about problems occurring in products during a user test.
Furthermore, it is investigated whether the problems found by a selection of children based on personality characteristics is representative for the problems found by other children. The remainder of this chapter is divided into six sections. First, the measures to determine the suitability of each child for participation in a user test and his/her personality characteristics are described, resulting in a set of hypotheses. Then, the set-up of an experiment to test these hypotheses will be described. The next section describes the data analysis process that was used in order to obtain the relevant measures. Subsequently, the results are presented, accompanied by a discussion of the representativeness of the detected problems. Finally, the generalisability of the results is discussed and conclusions are drawn.

5.2 User test outcome and personality characteristics

5.2.1 User test outcome

In order to compare how well different children are able to participate in a user test, some specific measures are needed. A first measure of how well a child is able to participate in a user test is the number of problems revealed by that child. Each product has a number of problems, which could be fixed to increase the quality of the product. During a user test, participants must help to identify these problems. A ‘good’ user test participant is one who can assist in finding a large proportion of these problems.

A second measure to determine the suitability of a child for participation in a user test is the ratio of problems indicated through user-initiated spoken output. As discussed in the introduction, problems in children’s products can be discovered during a user test by the observation of interaction with the product, by the user-initiated spoken output of the child, and by a combination of observation and user-initiated spoken output. Problems that are not indicated by the spoken output of a child must be based solely on observation of interaction with the product and are more likely to be missed by the evaluator. Furthermore, it is often much easier for an evaluator to determine the causes of detected problems when children give verbal comments. For example, if a child clicks randomly on a navigation-screen, the evaluator could reason that the child does not know what the purpose of the screen is, either because it was not explained properly, or because the child cannot distinguish between clickable and non-clickable elements. If, in addition, the child says: ‘Where do I have to click to go further?’ the evaluator will be more certain that the cause of the problem is that the
child does not recognise the clickable elements. The second measure is defined more precisely in equation (5.1).

\[ \text{Ratio verbally indicated problems(i)} = \frac{\# \text{ verbal problems(i)}}{\# \text{ all problems(i)}} \quad (5.1) \]

Where \( \# \text{ verbal problems(i)} \) is the number of problems indicated through user-initiated spoken output (possibly in combination with non-verbal behaviour) of child i, and \( \# \text{ all problems(i)} \) is the total number of problems found by the evaluators in the test session with child i.

5.2.2 Personality characteristics

To describe the personality characteristics of children, a set of validated and reliable measures is needed. These measures should be easy to obtain in order to function as a practical selection mechanism. Young children are not yet able to complete questionnaires, and they are not yet able to self-reflect. Therefore, an instrument for this age-group should be based on observations by parents or caretakers. The Blikvanger 5-13 (Elphick et al., 2002) is the only instrument in The Netherlands that describes non-pathological personality traits based on observations of parents or caretakers for children between five and 13 years old. Blikvanger 5-13 has shown to have an internal consistency of Cronbach’s \( \alpha \geq 0.80 \) for each of the main scales, and for all but three of the subscales. For a discussion of the convergent and divergent validity of the Blikvanger 5-13 see (Elphick et al., 2002). The Blikvanger 5-13 covers five main personality traits, called the Big Five, which are commonly used in many personality tests. These personality traits are Extraversion, Friendliness, Conscientiousness, Emotional stability, and Intelligence. These five main scales are divided into 16 subscales with 8 items each, resulting in a questionnaire of 128 items describing personality aspects that parents have to score on a five-point scale. The subscales for each of the five main scales are given in Table 5-1.

Table 5-1 Scales and subscales of the Blikvanger 5-13 in English and Dutch

<table>
<thead>
<tr>
<th>Main scale</th>
<th>Main scale in Dutch</th>
<th>Subscales</th>
<th>Subscales in Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td>Extraversie</td>
<td>Outgoingness</td>
<td>Toenadering zoeken</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Positive emotionality</td>
<td>Positieve emotionaliteit</td>
</tr>
</tbody>
</table>
The questionnaire results are entered into a software package that creates an individual profile for the child. For the five main scales and each of the subscales, a score is calculated which is visualised in two graphs. In these graphs the norm scores, based on the results of 106 mothers who completed the questionnaire, are also visualised. Furthermore, the grey area in these graphs represents the scores that fall within the interval of $-1$ and $+1$ standard deviations of the normal score. Examples of the two graphs belonging to one individual profile of a child are given in Figure 5-1 and Figure 5-2.
5 Personality characteristics as predictors of participant effectiveness

Normscores child compared to reference group
Normgroup: General average

Figure 5-1 Example of a visualised representation of scores on the five main scales as in Blikvanger report. The grey area in this graph represents the scores that fall within the interval of −1 and +1 standard deviations of the normal score.
Before the actual study about personality characteristics was performed a pilot study with seven children, each playing the games ‘Rainbow, The most beautiful fish in the ocean’ (MediaMix, 2002b) and ‘Roger Rabbit, Fun in the clouds’ (MindScape, 2003). The parents had filled in the Blikvanger questionnaire prior to the test. In this pilot study the children did not have to perform any specific tasks, they were asked to think aloud but were not prompted, and they could receive help from the facilitator (as described in Chapter 2). After analysing the user tests in the same way as described in Chapter 3 an exploratory analysis was performed to determine which main scales and sub scales could be successful predictors for the number of experienced problems and the ratio of verbalised problems.

Based on this analysis two main hypotheses were formulated. These hypotheses relate to the two measures discussed in section 5.2.1; the number of problems and the ratio of verbally indicated problems. The first hypothesis concerns the personality characteristic that influences the number of problems. The rationale behind this hypothesis is that curious children will try out more things and show more unpredictable behaviour, and will therefore encounter more aspects of the product that can cause problems. The first hypothesis is:

5.3 Experiment: Hypotheses

Figure 5-2 Example of a visualised representation of scores on the subscales in Blikvanger report (in Dutch). The grey area in this graph represents the scores that fall within the interval of −1 and +1 standard deviations of the normal score.
H1. There is a significant positive correlation between the score on Curiosity and the number of problems.

The second hypothesis concerns the combination of personality characteristics that influence the ratio of verbalised problems. The first part of the second hypothesis is that extravert children will be more inclined to seek contact with the facilitator by talking to him or her. This assumption is quite similar to the one made by Donker and Markopoulou (2001), who reasoned that extraversion might significantly affect the likelihood that children voice their thoughts about usability problems they encounter, and therefore tend to increase the number of found problems. Note, however, that extraversion in the present experiment is hypothesised to affect the ratio of verbalised problems, not simply the number of problems. The second part of this hypothesis is that children who score not very high on Friendliness will be more inclined to blame the product than themselves for their problems and will therefore make more comments about these problems to the test facilitator. The combination of these factors could be an indication of how much a child will actually talk about problems that occur during the test. This results in the following hypothesis:

H2. There is a significant correlation between the scores on Extraversion and Friendliness, and the proportion of the problems indicated through self-initiated spoken output. This correlation is positive for Extraversion and negative for Friendliness.

During a discussion of the available personality characteristics in the Blikvanger it was decided that one other characteristic could be a candidate to predict the percentage of verbalised problems. This third hypothesis is thus based on the definition of Autonomy in the Blikvanger and not on experience from the pilot tests: An autonomous child seldom asks for help. Therefore children who are less autonomous will ask for help more often, making them verbalise their problems in order to receive help from the facilitator. This results in the third hypothesis:

H3. There is a significant negative correlation between the score on Autonomy and the proportion of the problems indicated through self-initiated spoken output.

5.4 Exploring other predicting factors

The given hypotheses are based on earlier experiences with children during user tests in the lab. However, it may also be interesting to determine whether there are any overlooked personality characteristics that may be good predictors for the number of detected problems,
or the proportion of the problems indicated through self-initiated spoken output. Therefore, exploratory regression analyses will be performed on all gathered data.

### 5.5 Representativeness of problems after selection

It is important to determine whether the selection of children based on personality characteristics may inadvertently cause the detection of a non-representative subset of problems for the whole user group. It is especially important to check whether this selection of subjects, in this case children, would cause more serious problems to remain undetected (Law and Hvannberg, 2004) than can be expected when making a selection. For this purpose, all problems will be categorised according to two severity measures. These severity measures are Frequency severity and Impact severity. These measures are similar to many commonly used severity measures (Rubin, 1994, Nielsen, 1993). Subsequently, section 5.10.2 describes how many problems would have been found if only a small group of most promising children had been used and what the severity of these problems would have been. If the smaller group of most-promising children does not find less high- and medium-severe problems than can be expected from a group of this specific size, then the problems detected by this group are representative for the problems detected by other children.

### 5.6 Method

#### 5.6.1 Participants

To test the hypotheses an experiment was set up with 26 children of group three and four (grade one and two) of De Brembocht, an elementary school in Veldhoven, The Netherlands. This school is situated in a neighbourhood that is mainly inhabited by people who received higher education and earn more than minimum wage. All children were between five and seven years old (mean age was 84 months, SD=5.6 months), 9 girls and 17 boys. They were recruited by a letter to their parents asking for their cooperation. When the parents indicated consent they had to complete the Blikvanger 5-13 questionnaire prior to the test session. Of the 31 parents that were willing to let their child participate in the experiment one did not complete all the questions of the Blikvanger 5-13 questionnaire, and the data on this child were therefore discarded. One child suffered from Down’s syndrome. Therefore, she only participated in the evaluation but her data were not used in the analysis. Furthermore, the
test sessions of three other children were not videotaped correctly. Therefore the data on these three children were also discarded, resulting in 26 children in the experiment.

5.6.2 Test material

The 26 children in the experiment were asked to participate in a user test of a computer game called ‘Milo and the Magical Stones’ (MediaMix, 2002a). This game is intended for children between four and eight years old and is a good representative of software products for children of this age group. This game is an adventure game and it is a typical representative of software products for children between five and seven years old. It consists of ten sub games, two navigational screens, three story screens, one help-screen and one stop-screen. Children have to play all sub games in order to find magical stones for a group of mice. These stones are meant to keep all the mice warm during the winter and are spread through the game. Among these sub games there are e.g. motor-skill games, logical games, and musical games. Many problems were anticipated for children playing this game because even the adult researchers had some problems playing it. This would make the game suitable for the experiment.

5.6.3 Procedure

Each child was taken from the classroom separately for 30 minutes to perform a user test with the game. First the test facilitator explained the purpose of the test and instructed the child to try to talk aloud. The child could play the game as he or she liked without any specific tasks, the test facilitator did not remind the child to talk aloud during the test, and help was given as described in Chapter 2. After 25 minutes the facilitator would tell the child that the test session was over but that he/she could choose to continue playing the game for another five minutes or return to the class. If the child chose to continue playing, the session was stopped after 30 minutes in total. Each test session was videotaped, recording a split-screen shot of the face of the child and the onscreen actions.

5.7 Analysis

For detecting and coding the video data in order to detect problems, again Observer Pro (Noldus, 2002) was used, a software package for observational research. The coding scheme described in section 3.3 was used to code breakdown indications. Eight out of the 26 user tests were analysed with the Observer by two evaluators. To check the inter-coder reliability
for these two evaluators the any-two agreement measures were calculated for the results of the individual breakdown coding. The average any-two agreement measure for these eight analysed user tests was 0.73. The evaluators discussed all coding and clustering into problems to determine the final list of problems for each child.

### 5.8 Determining the percentage of verbally indicated problems

The ratio of verbally indicated problems for each child was determined by first counting the total number of problems written down in the problem reports, and subsequently by counting the number of verbally indicated problems. A verbally indicated problem was defined as a problem that is detected based on at least one breakdown indication that corresponds with a verbal comment in the transcription.

For example, if a child clicks the exit-button while trying to play the game (wrong action) and says “Oops, this is to quit the game!” (recognition), this problem is counted as a verbally indicated problem because it is indicated by two breakdown indications, of which one (recognition) has a corresponding verbalisation in the transcript (“Oops, this is to quit the game!”). In contrast, if another child just clicks the exit button (wrong action) and does not say anything, this problem is not counted as a verbally indicated problem, but just as a problem.

Finally, the number of verbally indicated problems was divided by the total number of problems to calculate the ratio of verbally indicated problems.

### 5.9 Determining representativeness

As described in section 5.4 it is important that the selection of children on the basis of some of their personality characteristics still ensures the detection of representative problems based on their user tests. It would be unsatisfactory when selection of most-promising children for the user tests would cause serious problems to remain undetected. To address the question of representativeness of the problems found by a selection of most-promising children, all problems will first be categorised according to the two severity measures described in Chapter 4: Frequency Severity and Impact Severity. However, a classification is now made to distinguish high severity, medium severity, and low severity problems. Subsequently, the problems that are missed by a group of most-promising children are discussed in terms of these classifications of both types of severity.
The classification of problems into different severity categories is performed by choosing sensible thresholds based on the context of the evaluated product. Nielsen (1993) defines only two levels for each type of severity, while Rubin (1994) defines four levels of severity. Woolrych and Cockton (2001) define three levels of severity for each type. For the classification in this experiment it was also decided to define three levels for both types of severity. Using two levels would be too crude a classification to see a difference between problems that can be overcome by the children themselves, with a little help, or with a lot of help. Four levels would be too fine grained, because the highest level of frequency severity would not contain a sufficient number of problems due to the fact that children could visit different parts of the game. The exact choices made for these levels are described in the next sub sections.

**Frequency Severity**

The three levels of frequency severity for this experiment were chosen by determining what the highest percentage of children experiencing a problem in this study was. The highest percentage was only 65%, which equals 17 children. However, there was only one problem experienced by so many children. The second highest severity was 57% or 15 children. This percentage of 57% was split into three equal categories to represent high, medium and low frequency severity. The Frequency Severity classification of each problem was thus determined as follows:

- High frequency severity: Problem was experienced by ≥38% of the children.
- Medium frequency severity: Problem was experienced by 20-37% of the children.
- Low frequency severity: Problem was experienced by ≤19% of the children.

**Impact Severity**

The three levels of Impact Severity were created by dividing the Overall Impact Severity which ranges from 1.0 to 3.0 as described in section 4.9.3, into three equal parts:

- High impact severity: 2.4< Overall Impact Severity ≤3.0
- Medium impact severity: 1.7< Overall Impact Severity ≤2.4
- Low impact severity: 1.0≤ Overall Impact Severity ≤1.7
5.10 Results

5.10.1 Hypotheses

Linear regression analyses were performed to test the hypotheses. Because all hypotheses were formulated as one-sided hypotheses a significance-level of 0.10 was chosen. This significance-level is based on the commonly accepted level of significance of 0.05 for testing two-sided hypotheses in behavioural studies. The first hypothesis asserts that:

H1. There is a significant positive correlation between the score on Curiosity and the number of problems.

The regression analysis revealed a significant effect for Curiosity on the number of problems ($df=25, F=5.864, R^2=0.196, p=0.023$).

The second hypothesis asserts that:

H2. There is a significant correlation between the scores on Extraversion and Friendliness and the proportion of the problems indicated through self-initiated spoken output. This correlation is positive for Extraversion and negative for Friendliness.

Linear regressions were first performed for each of the separate predictors, Extraversion, and Friendliness on the ratio of verbal problems. The analysis revealed no significant effect ($p>0.10$) for these separate predictors on the ratio of verbal problems. However, there was a significant effect for the combination of Extraversion and Friendliness ($df=25, F=4.971, R^2=0.302, p=0.016$) on the ratio of verbal problems in the expected directions (Extraversion positive, Friendliness negative).

The third hypothesis asserts that:

H3. There is a significant correlation between the scores on Autonomy and the proportion of the problems indicated through self-initiated spoken output.

The analysis revealed no significant effect for the score on Autonomy on the ratio of verbal problems ($F=0.138, p>0.10$).

5.10.2 Exploring other factors

Main scales

Explorative regression analyses were performed for all possible subsets of the main scales (Extraversion, Friendliness, Conscientiousness, Emotional stability and Intelligence) on the
two performance variables. The results for the number of problems are given in Table 5-2. The results for the proportion of problems indicated through self-initiated spoken output are given in Table 5-3. However, the results from these regression analyses should be treated with much more caution than the outcomes of the tested hypotheses because doing many analyses on the same data may lead to invalid conclusions. Regression analyses and especially related techniques such as path analysis are not credible without associated theories and hypotheses, otherwise artefacts of the data can be interpreted as results, which they are not.

Table 5-2 Regression analyses for each possible subset of main scale predictors for the number of problems. The first five columns indicate the variables present in the tested subset. The results are given in ascending order, beginning with one-predictor equations and concluding with the five-predictor equation. At each stage R²'s are given in descending order.

<table>
<thead>
<tr>
<th>Variables in model</th>
<th>F</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friendliness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional Stability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intelligence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x**</td>
<td>7.048</td>
<td>0.227</td>
</tr>
<tr>
<td>x</td>
<td>3.290</td>
<td>0.121</td>
</tr>
<tr>
<td>x</td>
<td>0.653</td>
<td>0.026</td>
</tr>
<tr>
<td>x</td>
<td>0.528</td>
<td>0.022</td>
</tr>
<tr>
<td>x</td>
<td>0.059</td>
<td>0.002</td>
</tr>
<tr>
<td>x</td>
<td>6.453</td>
<td>0.359</td>
</tr>
<tr>
<td>x</td>
<td>4.415</td>
<td>0.277</td>
</tr>
<tr>
<td>x</td>
<td>4.372</td>
<td>0.275</td>
</tr>
<tr>
<td>x</td>
<td>3.789</td>
<td>0.248</td>
</tr>
<tr>
<td>x</td>
<td>3.716</td>
<td>0.244</td>
</tr>
<tr>
<td>x</td>
<td>1.584</td>
<td>0.121</td>
</tr>
<tr>
<td>x</td>
<td>0.872</td>
<td>0.070</td>
</tr>
<tr>
<td>x</td>
<td>1.878</td>
<td>0.066</td>
</tr>
<tr>
<td>x</td>
<td>0.467</td>
<td>0.039</td>
</tr>
</tbody>
</table>
5 Personality characteristics as predictors of participant effectiveness

<table>
<thead>
<tr>
<th>Variables in model</th>
<th>Extraversion</th>
<th>Friendliness</th>
<th>Conscientiousness</th>
<th>Emotional Stability</th>
<th>Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>0.403</td>
</tr>
<tr>
<td>x^*</td>
<td>x^*</td>
<td>-</td>
<td>-</td>
<td>x^*</td>
<td>4.643</td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>x^*</td>
<td>4.558</td>
</tr>
<tr>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>x^*</td>
<td>4.430</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>x^*</td>
<td>x</td>
<td>x^*</td>
<td>3.202</td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x^*</td>
<td>2.976</td>
</tr>
<tr>
<td>-</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x^*</td>
<td>2.399</td>
</tr>
<tr>
<td>x</td>
<td>x^*</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>2.379</td>
</tr>
<tr>
<td>x^*</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>1.343</td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>0.744</td>
</tr>
<tr>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>4.431</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>3.695</td>
</tr>
<tr>
<td>x^*</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x^*</td>
<td>3.545</td>
</tr>
<tr>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x^*</td>
<td>4.870</td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x^*</td>
<td>3.434</td>
</tr>
<tr>
<td>x^*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>1.795</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x^*</td>
<td>3.436</td>
</tr>
</tbody>
</table>

*p<0.05, **p<0.01

The personality characteristic Intelligence seems to be a good predictor for the number of problems (df=25, F=7.048, R^2=0.277, p=0.014). This is not surprising because the hypothesised predictor Curiosity is one of the subscales of this main scale. Another good predictor in combination with Intelligence is Conscientiousness (df=25, F=6.453, R^2 =0.359, p=0.006). A high score on Intelligence combined with a low score on Conscientiousness gives
5 Personality characteristics as predictors of participant effectiveness

A high number of problems. Conscientiousness was not a hypothesised predictor but it makes sense that children who are curious and not very careful in what they try or how they try it because of a low conscientiousness, are likely to experience many problems.

Table 5-3 Regression analyses for each possible subset of main scale predictors for the proportion of problems indicated through self-initiated spoken output. The first five columns indicate the variables present in the tested subset. The results are given in ascending order, beginning with one-predictor equations and concluding with the five-predictor equation. At each stage $R^2$'s are given in descending order.

<table>
<thead>
<tr>
<th>Variables in model</th>
<th>Extraversion</th>
<th>Friendliness</th>
<th>Conscientiousness</th>
<th>Emotional Stability</th>
<th>Intelligence</th>
<th>$F$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.406</td>
<td>0.124</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.256</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>0.951</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>0.077</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>0.006</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>x*</td>
<td>x**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.971</td>
<td>0.302</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>1.838</td>
<td>0.138</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>1.637</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>1.645</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>1.087</td>
<td>0.086</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>0.742</td>
<td>0.061</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>0.691</td>
<td>0.057</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>0.635</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>0.510</td>
<td>0.042</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>0.037</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>x*</td>
<td>x**</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>3.276</td>
<td>0.309</td>
<td></td>
</tr>
<tr>
<td>x*</td>
<td>x*</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>3.222</td>
<td>0.305</td>
<td></td>
</tr>
</tbody>
</table>
5 Personality characteristics as predictors of participant effectiveness

<table>
<thead>
<tr>
<th>Variables in model</th>
<th>Extraversion</th>
<th>Friendliness</th>
<th>Conscientiousness</th>
<th>Emotional Stability</th>
<th>Intelligence</th>
<th>$F$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>x* x* - x -</td>
<td>3.189</td>
<td>0.303</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- x x x -</td>
<td>1.370</td>
<td>0.157</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- x x - x x</td>
<td>1.233</td>
<td>0.144</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x - x x x -</td>
<td>0.718</td>
<td>0.089</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x - x - x x</td>
<td>0.693</td>
<td>0.086</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x - - x x x</td>
<td>0.494</td>
<td>0.063</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x* x* - x x x</td>
<td>2.346</td>
<td>0.309</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x* x* x - x</td>
<td>2.352</td>
<td>0.309</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x* x* x x -</td>
<td>2.308</td>
<td>0.305</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- x x x x</td>
<td>1.005</td>
<td>0.161</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x - x x x x</td>
<td>0.515</td>
<td>0.089</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x x x x</td>
<td>1.794</td>
<td>0.310</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p<0.05, **p<0.01

Apart from the hypothesised main scale predictors Extraversion and Friendliness there does not seem to be an alternative predictor or set of predictors for the proportion of problems indicated through self-initiated spoken output.

**Sub scales**

For the 16 sub scales the number of all possible subsets is too large ($2^{16}=65536$ subsets) to perform all possible regression analyses as was done for the main scales. Therefore, stepwise, backward, and forward regressions are performed to determine whether there are any other
potential predictors for the number of problems and the proportion of verbalised problems than the ones hypothesised.

For the number of problems both the stepwise and the forward regression analysis indicates a combination of Dominance and Curiosity as set of predictors ($df=25$, $F=6.040$, $R^2=0.344$, $p=0.008$). Curiosity was the hypothesised predictor. Dominance is a new predictor which could be investigated further.

Backward regression indicates a much larger set of predictors, containing Self-Confidence, Outgoingness, Curiosity, Emotional stability, and Manageability ($df=25$, $F=5.011$, $R^2=0.556$, $p=0.004$). Curiosity is the only predictor that is present in all results; the additional predictors could be investigated further.

For the proportion of problems indicated through self-initiated spoken output both the stepwise and the forward regression analysis indicates a combination of Outgoingness and Altruism as set of predictors ($df=25$, $F=4.874$, $R^2=0.298$, $p=0.017$). Since Outgoingness is part of the main scale Extraversion and Altruism is part of the main scale Friendliness this is in agreement with the hypothesis. No additional predictors are indicated by these analyses.

Backward regression indicates a much larger set of predictors, containing Altruism, Outgoingness, Affection, Emotional stability, Manageability, Dominance, and Positive emotionality ($df=25$, $F=3.083$, $R^2=0.545$, $p=0.026$). Altruism and Outgoingness are the only predictors present in all analysis results; these additional predictors could be investigated further.

5.10.3 Representativeness

By determining whether a selection of the ‘most-promising’ children would have resulted in missing an unexpectedly high percentage of problems or many high-severity problems it can be decided whether the problems uncovered by these children are representative of the problems of this subgroup. However, when a smaller group is selected than the 26 children in the whole experiment, the expectations for the percentages of detected problems should also be lowered. The formula $1 - (1 - p)^n$, where $n$ is the number of test participants and $p$ is the average detection rate of all problems can be used to calculate what percentage of problems will be detected when using $n$ participants (Nielsen, 1994). Although Nielsen (1994) and Virzi (1992) have reported average detection rates $p$ as high as 0.42, detection rates are often much lower (Lewis, 1994, Bekker et al., 2004, Woolrych and Cockton, 2001) because they depend on the test set up, the test application and the test users chosen. To compare the actual
percentages of detected problems by a group of most-promising children for each severity category to the expected percentages for each severity category, the actual $p$-value should be determined based on the data gathered in this experiment. Based on the results of the 26 children it was calculated that the detection rate in this experiment was as small as 0.14. In practice, many evaluators use a problem detection of about 80% of all problems as sufficient. Although aiming at detecting 80% of all problems may not be good practice (Cockton and Woolrych, 2001) this percentage will only be used to show that a selection of children will not detect unexpectedly low percentages of problems of the three types of severity.

By using the formula $1 - (1 - p)^n$ with $p=0.14$ it was calculated that 11 children should have been selected to find 80% of all problems. The selection of these 11 children should weigh both beneficial factors, Curiosity as well as Extraversion/Friendliness. Therefore, a ranking of the children was made. First the children were ranked according to Curiosity. Subsequently the children were ranked according to the Extraversion/Friendliness combination score. Based on the regression equation, the combination score was calculated by subtracting the score for Friendliness from the score for Extraversion. Because the ranking on Curiosity and the ranking on Extraversion/Friendliness combination were decided to be equally important they were simply added up to come to one overall ranking. For example, if a child ranked 11th on Curiosity and 12th on Extraversion/Friendliness the overall rank for this child would be 23. The 11 children with the highest overall rank were chosen to represent the ‘most-promising’ group.

**Number of found problems**

The first measure of representativeness is the actual percentage of problems found by this group of most-promising children, compared to the expected percentage of found problems. The expected percentage is 80%. All 26 children together found 109 problems. The group of 11 most-promising children would have found 82 problems, which is 75% of all problems. This is not far from the expected percentage.

**Severity of found problems**

For each severity category within the two types of severity, Frequency Severity and Impact Severity, the group of most-promising children is again expected to find 80% of the problems. Of the 109 problems found by all children, 11 were classified as high frequency severity problems, 20 as medium frequency severity problems, and 78 as low frequency severity
problems based on the Frequency Severity classification. The 11 most-promising children would have found 100% of the high frequency severity problems, 100% of the medium frequency severity problems, and 65% of the low frequency severity problems. Of all 109 problems, six were classified as high impact severity problems, 18 as medium impact severity problems, and 85 as low impact severity problems based on the Impact Severity classification. The 11 most-promising children would have found 83% of the high impact severity problems, 88% of the medium impact severity problems, and 72% of the low impact severity problems.

The data show that for each type of severity the 11 most-promising children find at least as many high and medium severity problems as expected. Therefore, it can be concluded that the selection of most-promising children based on the personality characteristics Extraversion, Friendliness, and Curiosity, would not have seriously reduced the representativeness of the detected problems for randomly chosen group of 11 children.

5.10.4 Example of the effects of selection

To give an example of the effects that choosing a group of most-promising children can have on the results of a user test, the results of this group will be compared to those of a group of least-promising children. The group of least-promising children consists of the 11 children that had the lowest overall ranking as described in section 5.10.3.

The least-promising group of children would have found 76 problems, compared to 82 for the most-promising group. In the least-promising group only 28 problems would have been indicated verbally by at least one child while in the most-promising group 43 problems would have indicated been verbally by at least one child. Finally, of the 44 problems that would have been found by both groups, the average number of children that would have found each problem is significantly lower than in the least-promising group. The comparison of the results of both groups of children is given in Table 5-4.

Table 5-4 Comparison of the numbers of detected problems, and verbalised problems, and the average numbers of children detecting a problem for the group of most-promising and least-promising children.

<table>
<thead>
<tr>
<th></th>
<th>Group of most-promising children</th>
<th>Group of least-promising children</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of found problems</td>
<td>82</td>
<td>76</td>
</tr>
</tbody>
</table>
Personality characteristics as predictors of participant effectiveness

<table>
<thead>
<tr>
<th></th>
<th>Group of most-promising children</th>
<th>Group of least-promising children</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of verbalised problems</td>
<td>43</td>
<td>28</td>
</tr>
<tr>
<td>Average number of children finding a problem (^a)</td>
<td>3.5</td>
<td>2.8*</td>
</tr>
</tbody>
</table>

\(^a\) Over 44 problems found by both groups

\(^*p<0.001,\) one-tailed

Regarding Frequency Severity, the group of least-promising children would have found 100% of both the medium and high frequency severity problems. Of the 78 low frequency severity problems, they would have found 58%. The comparison of these results to the results of the most-promising group of children and the whole group of 26 children is given in Table 5-5.

Table 5-5 Numbers of high, medium, and low frequency severity problems that would have been found when a selection of the 11 least-promising children would have been made, compared to the numbers of problems found by the 11 most-promising children and to the numbers of problems found by all 26 children.

<table>
<thead>
<tr>
<th>Frequency severity category</th>
<th># Problems found by all 26 children</th>
<th># Problems found by 11 most-promising children</th>
<th># Problems found by 11 least-promising children</th>
</tr>
</thead>
<tbody>
<tr>
<td>High severity</td>
<td>11</td>
<td>11 (= 100%)</td>
<td>11 (= 100%)</td>
</tr>
<tr>
<td>Medium severity</td>
<td>20</td>
<td>20 (= 100%)</td>
<td>20 (= 100%)</td>
</tr>
<tr>
<td>Low severity</td>
<td>78</td>
<td>51 (= 63%)</td>
<td>45 (= 58%)</td>
</tr>
</tbody>
</table>

Both groups of children find all high and medium frequency severity problems and there is only a small difference in the number of low frequency severity problems. This is not surprising since medium and high frequency severity problems were found by at least 19% of the 26 children (= 5 children) in the experiment. With a group of 11 children it can be expected that all problems in these categories are found.

Regarding Impact Severity, the group of least-promising children would have found 83% of both the high impact severity problems. Of the 18 average impact severity problems, they would have only found 61%. Of the low impact severity problems they would have found 71%. The comparison of these results to the results of the most-promising group of children and the whole group of 26 children is given in Table 5-6.
Table 5-6 Numbers of high, medium, and low impact severity problems that would have been found when a selection of the 11 least-promising children would have been made, compared to the numbers of problems found by the 11 most-promising children and to the numbers of problems found by all 26 children.

<table>
<thead>
<tr>
<th>Impact category</th>
<th>Problems found by all 26 children</th>
<th>Problems found by 11 most-promising children</th>
<th>Problems found by 11 least-promising children</th>
</tr>
</thead>
<tbody>
<tr>
<td>High severity</td>
<td>6</td>
<td>5 ( = 83%)</td>
<td>5 ( = 83%)</td>
</tr>
<tr>
<td>Medium severity</td>
<td>18</td>
<td>16 ( = 88%)</td>
<td>11 ( = 61%)</td>
</tr>
<tr>
<td>Low severity</td>
<td>85</td>
<td>61 ( = 72%)</td>
<td>60 ( = 71%)</td>
</tr>
</tbody>
</table>

The only difference between the two groups of children in Table 5-6 is the number of medium impact severity problems. An explanation for this difference is that medium impact severity problems are typically problems that can be solved by the facilitator giving some help (because of the way Impact Severity is determined). Because children only received help when they repeatedly asked for it, the chance of detecting such problems is higher for children who are not reluctant to talk about problems and ask help from the facilitator.

The comparison of the results of the group of most-promising children and the group of least-promising children shows three important differences. The first difference is that the group of most-promising children indicated more problems verbally. The second difference is that in the group of most-promising children a certain problem was detected by more children than in the least-promising group. The third difference is that the group of most-promising children found more medium impact severity problems than the group of least-promising children. Thus, this example clearly illustrates how the selection of a group of most-promising children instead of least-promising children, based on the personality characteristics Curiosity, Friendliness and Extraversion, would have been beneficial for the results of this user test.

### 5.11 Discussion

#### 5.11.1 Related research

Other researchers have tried to find good predictors for the effectiveness of children participants in user tests. However, a study by Donker and Markopoulos (2001), which had a
very similar research question as the one presented in this experiment, did not find the effect of extraversion and verbal competence on the number of detected problems with different
evaluation methods to be significant. However, the experiment in this chapter indicates that
extraversion should not be considered as indicator of the number of problems, but merely as
an indicator of whether the child will give any verbal comments once a problem arises.
Furthermore, this study indicates that extraversion should not be considered without also
considering friendliness.
This study also gives experimental foundation for the common practice of using children who
are characterised as not shy for evaluation purposes, like for example Hanna et al. (2004) did
for their evaluation of game concepts. However, this study indicates that other personality
characteristics like friendliness and curiosity are also of influence.

5.11.2 Generalisability
Care should be taken in extrapolating the concepts of Curiosity, Extraversion, and
Friendliness to other groups of children and different languages. The Blikvanger has been
normalised on Dutch children and parents and its concepts may not be directly translatable.
Few validated personality tests for children exist outside the Dutch language area. In order to
select non-Dutch speaking children to participate in an evaluation it may be necessary to
develop a validated personality test based on the Blikvanger.
Regarding the generalisability to other games there is an important indication that the results
will also hold for different games. In the pilot study with seven children on which the
hypotheses were based, the children were ordered according to their obtained ratio of verbal
problems. It appeared that the ordering remained almost the same over the two games. This
means that children who performed well on testing one game generally also performed well
on testing the other game and vice versa. This informal result indicates that it is likely that
the specific game is not of influence in how well a child performs in the user test.
For the selection of adult test users the results could still hold if the same protocol of
voluntary talking aloud would be used. However, for adults usually the strict think-aloud
protocol is followed according to which they are reminded to keep on talking. Because the
verbalisations are not self-initiated in this situation it could be that the personality
characteristics of the adults are of less importance. On the other hand, it is not unthinkable
that these personality characteristics also influence the ability to perform standard thinking-
aloud.
The trend found in this experiment does not necessarily restrict itself to games. The same reasoning about the willingness to explore and to communicate with the facilitator and the tendency to blame the product instead of oneself that lies behind the hypotheses would probably hold for other products.

5.11.3 Practical advice

As was shown in section 5.10.2, selecting children based on the personality characteristics, Extraversion, Friendliness, and Curiosity, does result in finding almost all high and medium severity problems. The advice for evaluation practitioners would be to make a selection of children based on the personality characteristics Extraversion, Friendliness and Curiosity. Preferably these children would find both a large number of problems due to Curiosity, and provide self-initiated spoken output for a large proportion of these problems due to high Extraversion and low Friendliness. However, these personality characteristics do not necessarily occur together; a child can score low on Curiosity but high on Extraversion/Friendliness, or vice versa. Therefore, the practitioner has to decide what is more important, finding many problems or getting more information about each problem. According to this decision the practitioner could select mainly on Curiosity or mainly on the combination of Extraversion and Friendliness.

5.12 Conclusions

The experiment described in this chapter examined whether personality characteristics of children can predict which children will make good test participants, i.e. find many problems and give verbal information about these problems. The results of this experiment showed that children who score high on Curiosity according to the Blikvanger reveal the highest number of problems. An explorative analysis suggests that a low score on the main scale Conscientiousness or a low score on the sub scale Dominance may also be good additional predictors for the number of problems, but this should be investigated further. The experiment also showed that children who score high on Extraversion but low on Friendliness according to the Blikvanger indicate the highest percentage of problems through self-initiated spoken output. Finally, children who score low on Autonomy according to the Blikvanger do not necessarily indicate a high percentage of problems through self-initiated spoken output. By choosing children based on these personality characteristics it is still possible to detect problems that are representative for the problems other children may
experience with the game. In order to perform effective evaluations with children, the selection of children may be a good instrument.
6 The Problem Identification Picture Cards method

In this chapter the development and assessment of a new evaluation method, called the Problem Identification Picture Cards method is described. This method prompts young children to express both usability and fun problems while playing a computer game. The method combines the traditional thinking-aloud method with picture cards that children can place in a box to indicate that there is a certain type of problem. An experiment to assess this method shows that children express more problems (verbally, or with a picture card, or with a combination of a picture card and a verbalisation) with the Problem Identification Picture Cards method than without this method (in which they can only indicate problems verbally). Children do not just replace verbalisations by using the provided picture cards and most children prefer to use the Problem Identification Picture Cards method during the test instead of standard thinking-aloud. The Problem Identification Picture Cards method can be a good instrument to increase the amount of information expressed by young children during an evaluation.

---

5 This chapter is based on the following position paper: (Barendregt and Bekker, 2005c)
6.1 Thinking-aloud with children

As described in Chapter 3 the commonly used ‘thinking aloud’ technique for discovering problems in a design (Nielsen, 1993) has the disadvantage that young children can have difficulty verbalizing their thoughts (Boren and Ramey, 2000). Because they often forget to think aloud, they need to be prompted to keep talking. However, prompting could result in children mentioning problems in order to please the experimenter, leading to non-problems being reported (Donker and Reitsma, 2004, Nisbett and Wilson, 1977). Therefore, the experiments in the previous chapters, as well as experiments by other researchers (Donker and Reitsma, 2004) relied on a combination of self-initiated spoken output complemented with observations of children’s behaviour. In the remainder of this chapter the self-initiated spoken output will be referred to as the results of the thinking-aloud method. Unfortunately, the amount of self-initiated spoken output in the thinking-aloud method is often limited. For example, in the experiment in Chapter 5 children on average only verbalised 31% of the problems that they encountered, and in Donker and Reitsma’s study (2004) only 28 out of 70 children made any remarks at all. Still, verbalisations or other clear signs of the child are very valuable because they may indicate problems that are likely to go undetected when relying on observations alone. For example, when a child thinks something is strange or silly, this is often difficult to detect unless the child says something about it. Furthermore, when an observable problem is accompanied by verbalisations or other explicit indications of a problem, the number of breakdown indications per problem increases, making it more likely that a problem will be detected by multiple evaluators (Vermeeren et al., 2002). Therefore, the reliability of a method that encourages children to express their thoughts while playing with the game will be higher.

In this chapter a new method that could help children express more of their thoughts than the thinking-aloud method is described and evaluated. First, the development and rationale of this new method is described. Subsequently, an experiment to test whether this method really encourages children to express more problems explicitly than the thinking-aloud method is described.

6.2 Development of the method

The first attempt to develop a new method to make children express more problems assumed that children might be too shy to verbalise their thoughts in front of an unfamiliar facilitator.
Based on literature about interviewing strategies in child assessment (Kanfer et al., 1983) it was hypothesised that children may talk more to someone they feel closer to than to the adult facilitator. The description of the Berkeley Puppet Interview method to assess children’s self-perception by Measelle et al. (1998) gave rise to the idea to equip the facilitator with a hand puppet. The hand puppet would try to build rapport with the child, and hopes were that children would try to engage the puppet in the game by talking to it about the game. In several pilot tests with children this idea was investigated with a cute hand puppet representing a fox.

![Figure 6-1 The fox hand puppet in the pilot study.](image)

However, it appeared that the method was hard to apply and would probably not give the expected results. There were several reasons for this failure:

- To give children the feeling that the hand puppet is real and engage them in a conversation the facilitator must be a rather good puppeteer. This makes it less suitable as a general method for facilitators.

- The children actually appeared to be very comfortable with the facilitator. Therefore they kept addressing the facilitator even when the hand puppet was present. The conversational situation thus contained three participants, which made it complex and unnatural for the facilitator to keep up.
On the advice of a play therapist it was decided to develop a method with picture cards that children can place in a box use to express different kinds of problems either verbally or non-verbally. There are several reasons why these picture cards would help children to express more problems explicitly than when the facilitator just asks the child to verbalise as much as possible about anything:

1. During the introduction the facilitator can use the picture cards to explain not only verbally but also visually what kind of information he/she is interested in. This combination of auditory and visual information adheres to the principles of multiple resources and redundancy gain (Wickens et al., 2004) and may make it easier for children to understand the explanation.

2. During the test the picture cards serve as memory aids for the things the evaluator is interested in, thereby putting ‘knowledge in the world’ (Norman, 1998) instead of ‘in the head’ and thus relying less on long-term memory.

3. Some children are able to verbalise what they think or feel, while others may be less verbally capable. With the picture cards method less verbally capable children can express themselves explicitly without having to verbalise. This is a similar approach as several interviewing techniques for young children (Measelle et al., 1998, Greca, 1983).

6.3 Choosing the pictures

Based on the problem types defined in the two taxonomies in Chapter 4, and the kinds of verbalisations made by children for each of these problem types in the previous studies, it was determined what concepts would represent their feelings when they encountered one of these problems. For example, when children verbalised a usability problem they often used phrases like “I don’t understand this” or “I don’t know what to do now”. When children encountered a problem related to too high a challenge level, they used a phrase like “This is too difficult!” These concepts show some overlap with items of the Fun-questionnaire for Kids (Stienstra and Hoonhout, 2002). This is not surprising since their questionnaire is also based on the work of Malone and Lepper (1987). However, the Fun-questionnaire for Kids does not include questions about usability problems. Furthermore, it is aimed at the overall assessment of a product instead of the feelings a child may have when a problem occurs.
Usability

- To be able to use a game, a child first needs to perceive the information given by the game. When a child encounters a perception problem he/she may say it is difficult to hear or see something clearly.

- When a child encounters a usability problem on the sensorimotor level of regulation (see section 4.3.1) he/she may find it difficult to use the mouse in order to click objects.

- When a child encounters a usability problem on another level than the sensorimotor level of regulation (see section 4.3.1) he/she may not understand what to do, or what has happened.

- When a child encounters an inefficiency he/she may think it takes too long.

Fun

- When a child encounters a fantasy that is aimed at older children he/she may find it scary.

- When a child encounters a fantasy that is aimed at younger children he/she may find it childish.

- When a child encounters a fantasy that is incongruent with the story or with his/her experiences he/she may find it silly or strange.

- When a child experiences a problem related to a too high challenge level he/she may find it too difficult.

- When a child experiences a problem related to a too low challenge level he/she may find it boring.

- When a child experiences a control problem he/she may think it takes too long.

- When a child experiences a curiosity problem he/she may find it boring.
While some concepts are the same for different problem types it was reasoned that the available context of the game would help to determine the meaning. To make clear to the children that the evaluation of a game is of course also about fun, one last concept ‘Fun’ was added.

In the first version of the picture cards, small icons were chosen from different online libraries to represent the different concepts. These icons were glued to wooden cards of about 2 x 2 cm. This first version was tested with two children in their home. Although the children did put some pictures in the box they had trouble picking up the small cards. Therefore it was decided to make bigger cards of about 4 x 4 cm. For these bigger cards clearer pictures were selected to represent the concepts. The final pictures were chosen from two on-line picture libraries that are also recommended for the PECS (Picture Exchange Communication System)-method. This PECS-method developed by Bondy and Frost (1994) is used to teach non-verbal autistic children to express themselves by exchanging picture cards. The libraries that were used to select pictures from are:


The pictures chosen for the Problem Identification Picture Cards method are given in Figure 6-2.

![Figure 6-2 From left to right, top to bottom, the pictures used for Boring, Don’t know/understand, Fun, Difficult, This takes too long, Childish, Silly and Scary, respectively.](image)

Each picture was glued to both sides of a wooden card. A wooden box with eight compartments was created in which children could place one of the cards when they encountered a problem that they wanted to express to the evaluator (see Figure 6-3).
6.4 The Problem Identification Picture Cards method

When using the Problem Identification Picture Cards (PIPC) method, children get an explanation of each picture and the kind of situation for which they can use it before the test session. During the test, the box and numerous picture cards for each problem category are placed on the table next to the computer on which the game is played. Children can place as many picture cards in the box as they like. The children can always ask for an explanation of a card if they happen to forget it. It does not matter whether they use the correct picture card for a particular problem. If the facilitator does not understand why a certain card is used he/she can ask the child for an explanation. Finally, the behaviour of the child with the game together with the picture cards is used to do the actual analysis of the test session.

6.5 Experiment: Evaluation of the PIPC method

The aim of the method was that children would express more problems, either verbally and/or using the picture cards than when they would just have been asked to verbalise as much as possible. To test whether the PIPC method would serve this aim, an experiment was set up to compare the two methods. The hypotheses concerning the differences between the PIPC method and the method solely relying on self-initiated spoken output are discussed in the next subsections.
6 The Problem Identification Picture Cards method

6.5.1 Hypothesis 1

Each picture card shows one of the pictures of Figure 6-2. These pictures represent the feelings children may have when they experience a problem (except for the Fun picture, which expresses enjoyment). Through the use of the picture cards children will probably have a clearer understanding of the feelings that they can communicate to the facilitator that indicate a problem. Furthermore, the picture cards may serve as a visual reminder of these feelings. Finally, children who are not so verbally capable can also express their feelings nonverbally by using a picture card. Therefore, the first hypothesis is that children will express more problems when they use the symbol cards method than when they have only been asked to verbalise as much as possible.

To test this hypothesis regression analyses will be performed to decide whether the difference in expressed problems between the methods can be explained by the game with which the methods are used, or by the order in which the methods are used, or on a combination of order and game. If this is not the case, a Wilcoxon signed ranks test will be performed to determine whether there is a significant positive effect of the PIPC method in the numbers of expressed problems.

6.5.2 Hypothesis 2

Although the picture cards give a clear indication of a problem, verbalisations can also give valuable information to an evaluator. Therefore the picture cards should be an addition to thinking-aloud; children should not just substitute verbal indications of problems with picture cards. The hypothesis is that this is not the case; the number of verbalised problems will not be lower for the PIPC method than for the thinking-aloud method.

To test this hypothesis a Wilcoxon signed ranks test will be performed on the numbers of verbalised problems with both methods.

6.5.3 Hypothesis 3

It is not always easy to find children who are willing and able to participate in a user test. Therefore, a user test should be a pleasurable experience to the children who participate so they like to participate again. The hypothesis is that children will not like the PIPC method less than the usual think-aloud method.
6.6 Method

6.6.1 Test participants

To test the hypotheses an experiment was set up with 23 children of group two (second year kindergarten) of the Wethouder van Eupen school, an elementary school in Eindhoven, The Netherlands. This school is situated in a neighbourhood that is mainly inhabited by people who received higher education and earn more than minimum wage. All children were five or six years old, twelve girls and eleven boys. They were recruited by a letter to the parents asking for their cooperation.

6.6.2 Experimental set-up

The results of the experiment described in Chapter 5 indicate that there are large differences between the percentages of problems children will verbalise depending on their personality characteristics. Therefore it was decided that in an experiment to test the effect of the PIPC method each child should perform each method.

Because the experiment described in Chapter 4 showed that the types of problems that children experience change when they become more experienced with a game, it was decided that children should play a different game for each method. These games should be of similar difficulty but different in the types of sub games that can be played.

It can be expected that children will learn from performing the first method and will thus perform better on the second method. To compensate for the order in which the children used the different methods each method should be used equally often as the first and as the second method.

Because the games should be different in the types of sub games that can be played it is not expected that the children will learn how to play the second game from playing the first game.

Altogether there were four different conditions and 23 children in the experiment. The children were randomly assigned to one of the conditions:

Table 6-1 Description of the four conditions in the experiment and the number of children in each condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Part 1 of the test</th>
<th>Part 2 of the test</th>
<th>Nr of children</th>
</tr>
</thead>
</table>
6 The Problem Identification Picture Cards method

<table>
<thead>
<tr>
<th>Condition</th>
<th>Part 1 of the test</th>
<th>Part 2 of the test</th>
<th>Nr of children</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>PIPC for game 2</td>
<td>Thinking aloud for game 1</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Thinking aloud for game 1</td>
<td>PIPC for game 2</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>Thinking aloud for game 2</td>
<td>PIPC for game 1</td>
<td>6</td>
</tr>
</tbody>
</table>

6.6.3 Test material

The 23 children in the experiment were asked to participate in a user test of two computer games ‘Milo and the red fruit’ (MediaMix, 2004b), and ‘Little Polar Bear, Do you know the way?’ (MediaMix, 2004a). These games are intended for children between four and eight years old and are good representatives of software products for children of the chosen age group of children between five and seven years old. They were put to market at the time of the experiment; therefore children would probably be unfamiliar with them. Furthermore, large numbers of problems were anticipated for children playing these games alone because even the adult researchers had some problems playing them. This would make the games quite suitable for the experiment.

6.6.4 Procedure

Each individual child was taken from the classroom for 50 minutes to perform two user test sessions; one for each method with a different game. First the test facilitator explained the general purpose of the test session, the procedure for the first method: either the think aloud condition or the picture cards method. The child then played the game for 15 minutes. As a training session the facilitator prompted the child extensively to talk aloud and/or use the cards during the first five minutes. During the subsequent 10 minutes the child could play the game as he or she liked without any specific tasks. When a child asked for help the first time, the test facilitator would only encourage the child to keep on trying. The second time a child asked for help the test facilitator would give a hint and only after the third time a child asked for help the facilitator would explain the solution in detail. After finishing the first test session, the child would get a short break of at most five minutes in which the facilitator started up the next game. After that the facilitator explained the next method for five minutes, and then prompted the child while playing the game extensively for five minutes. Finally the child played the second game with the next method alone for ten minutes. Each
test session was videotaped, recording a split-screen shot of the face of the child and the on-screen actions. A graphical representation of the procedure is given in Figure 6-4.

![Temporal representation of the test procedure](image)

Figure 6-4 Temporal representation of the test procedure (time in minutes).

At the end of the test session the child was asked to fill in a very short questionnaire. In this questionnaire the child had to mark with a cross which game he/she preferred and whether he/she preferred to do another evaluation in the future with or without the picture cards. The order of the possible answers was randomly changed to ensure that a preference for one of the games or with/without the picture cards was not due to the presentation of the answers. The questionnaire is given in Appendix 2.

### 6.7 Analysis

For each child the recorded video material was used to transcribe the protocols of both conditions for the ten minutes that they played the game without much interference of the facilitator (the two light grey boxes in Figure 6.4). For the picture cards it was also noted in the protocol when a child placed a picture card in one of the boxes (see example in Table 6-2). These protocols were used to count the number of unique problems (meaning that if a child experienced the same problem more than once, there was still only one problem counted) that were indicated verbally, with a picture card, or with a combination of a verbalisation and a picture card.

Table 6-2 Example of a transcribed protocol with problem count.

<table>
<thead>
<tr>
<th>Problem count</th>
<th>Verbalisation or picture card of the child</th>
<th>Facilitator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(places picture ‘don’t understand’ in box)</td>
<td>Yes, do you want me to help you?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>
A second evaluator checked these numbers by looking at the protocols and asking critical questions about why certain verbalisations were or were not taken into account, and whether certain verbalisations should be grouped or split. This review led to some minor changes in the final problem counts; for the PIPC method two problems expressed by one child were combined into one, and for four children a verbalisation was no longer counted as a problem, for the thinking-aloud method one verbalisation was removed as a problem and one was added.

6.8 Results

The results of the analysis of all protocols are given in Table 6-3. For testing the hypotheses in this experiment a significance-level of 0.10 was chosen because they are formulated as one-sided hypotheses, making a significance-level of 0.05, which is commonly used for two-sided tests, too restricting.
6 The Problem Identification Picture Cards method

Table 6-3 Numbers of expressed problems per child during the test with the think-aloud method and the PIPC method, and the number of verbalised problems during the PIPC method.

<table>
<thead>
<tr>
<th>Child</th>
<th>Expressed problems thinking aloud</th>
<th>Expressed problems PIPC</th>
<th>Verbalised problems PIPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>18</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

H1. None of the regression analyses for the games, the order of the methods, or the combination of these two factors on the difference between the numbers of expressed problems with both methods was significant \((p>0.10)\). The Wilcoxon signed ranks test showed that there was a significant positive difference between the number of problems expressed with the PIPC method and the thinking-aloud method \((Z=-2.024\) based on negative ranks, \(p<0.05\)).
H2. The Wilcoxon signed ranks test showed that there was no significant difference \( (p>0.05) \) between the number of verbalised problems with the PIPC method and the number of verbalised problems with the thinking-aloud method.

For the third hypothesis a Chi-Square test was performed on the expected and actual numbers of children who liked to perform another test with or without the PIPC method were compared:

H3. The number of children who would rather perform another user test with a new game with the PIPC method (14 of the 23 children=61\%) was significantly higher than the number of children who would rather perform another user test with a new game with the thinking-aloud method (9 out of the 23 children=39\%).

The lowest number of cards used by a single child was 0. The highest number of cards used by a single child was 9. A histogram of the numbers of children using any number of cards is given in Figure 6-5.
The choice frequencies of the picture cards for all children are given in Table 6-4.

Table 6-4 Number of times each picture card type was used by all children together.

<table>
<thead>
<tr>
<th>Card type</th>
<th>Number of times chosen (n=23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Don’t know/understand</td>
<td>19</td>
</tr>
<tr>
<td>This takes too long</td>
<td>8</td>
</tr>
<tr>
<td>Fun</td>
<td>5</td>
</tr>
<tr>
<td>Difficult</td>
<td>4</td>
</tr>
<tr>
<td>Silly</td>
<td>3</td>
</tr>
</tbody>
</table>
A Chi-Square test showed that the Don’t know/understand picture card was used significantly more often \((p<0.0001)\) than the other cards. This is not surprising since this card refers mainly to usability problems. The number of usability problems is usually much higher than the number of fun problems, as was also found in Chapter 4.

### 6.9 Discussion

#### 6.9.1 Games

The PIPC method was tested with two different adventure games. However, it is unclear whether the method also works for other game genres. Especially with very fast-paced games, children may not be able to give attention to the picture cards during the user test. An example of such a game for children in the chosen age group is ‘Freddi Fish, Silly Maze’ (Transposia, 2002). Further research is needed to determine which games could be evaluated with this method.

#### 6.9.2 Procedure

To minimise the intrusion on the normal routine in the school classes, the experiment had to be restricted to one test per child within one week. Furthermore, because young children have short attention spans, the maximum testing time had to be about 30 minutes to one hour (Hanna et al., 1997). Altogether this meant that the children had to perform the two test sessions in a short time. Consequently, the (training) time in which the facilitator could prompt for verbalisations or picture cards was also very short (five minutes). In these five minutes children did not encounter many problems for which the facilitator had the opportunity to prompt, so the children were not very well trained with each method. When performing an evaluation of one game, practitioners will usually have more time to train the test participants. This holds for both the PIPC method and the thinking-aloud method. Because no detrimental effects of the PIPC method on the number of verbalisations

<table>
<thead>
<tr>
<th>Card type</th>
<th>Number of times chosen (n=23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boring</td>
<td>3</td>
</tr>
<tr>
<td>Scary</td>
<td>1</td>
</tr>
<tr>
<td>Childish</td>
<td>0</td>
</tr>
</tbody>
</table>
was found it is likely that the PIPC method will still give a higher number of expressed problems than the standard thinking-aloud method, even with better training. However, the effect of a longer training time on the number of expressed problems should be examined further.

6.9.3 Possible improvements

In this study, the picture cards in their present form were an effective addition to the thinking-aloud method that relies solely on self-initiated spoken output. Still, several changes to the picture cards method are possible, but it has to be tested whether they would really be improvements.

Firstly, because the box in which to put the cards has to be put alongside the computer, children have to shift their attention from the screen to the box to place a picture card in one of the compartments. When the game initiates the interaction, for example when explaining something, children have to divide their attention between the game and the picture cards. Shifting attention from one display location to another requires effort (Wickens et al., 2004). Placing the pictures within closer proximity to the computer screen may make it easier for children to use the pictures in combination with the game.

Secondly, it was striking that only few children used a picture card without verbalizing anything. It seemed that the picture cards functioned much more as an aid to remember the things of interest than as an aid to help children who have difficulty verbalizing express their thoughts in a non-verbal way. This impression was also corroborated by the fact that some children just looked at the picture cards and then started verbalizing their thoughts. Maybe it is therefore not even necessary to ask children to put a picture in the box, but just to point to it.

Thirdly, the pictures used from the PECS-libraries were chosen because they were thought to express the feelings children would have when they encountered the different types of problems. The actual words associated with the pictures in the libraries were not always the same as the feelings or thoughts they had to represent. For example, the ‘jack-in-the-box’ picture was used to express ‘this is silly’. It is uncertain whether the pictures used were the best pictures for the different types of problems children can encounter when playing a game. However, children were not obliged to be able to remember the meaning of the pictures perfectly. When they forgot the meaning of a picture they could ask the facilitator. Therefore it was concluded that the pictures used were sufficient to remind the children of the concepts,
even when they were not perfect. Further research is needed to determine whether other pictures may be superior in expressing these concepts better.

Finally, it is possible that some of the concepts depicted by the pictures are superfluous, or that additional pictures are necessary. For example, the picture cards with ‘scary’ and ‘childish’ were almost never used. Therefore, further research is needed to determine the optimal set of pictures for the cards.

6.9.4 An unpredicted benefit of the PIPC method

One of the main advantages of the PIPC method that was not anticipated was the fact that it was much easier for the facilitator keep the attention of the children when explaining what the children were supposed to do. Although the facilitator tried to explain this in both conditions it was clear that many children could not keep their attention when the explanation was done only verbally. When using the picture cards it was much easier for the facilitator to explain the purpose of the test in a playful way by making the children guess the meaning of a certain picture and talk about it. Therefore the children could direct their attention to the explanation of the test situation while in the verbal condition their eyes were often drawn towards other things in the room. A post-hoc Wilcoxon signed ranks test showed that with the PIPC method children used the concepts explained with picture cards significantly more often than the same concepts explained verbally for thinking-aloud ($Z$=-3.26, $p<0.01$).

6.10 Conclusions

The problem identifying picture cards are a good addition to the thinking-aloud method based on self-initiated spoken output. When children can use these picture cards in addition to thinking-aloud they express more problems than with standard thinking-aloud. Children do not just replace verbalisations by picture cards without any verbalisations. The majority of the children like participating in a user test with the picture cards even better than without the picture cards. Whether other versions of the picture cards method (with more or fewer pictures, different pictures, and placement of the pictures closer to the computer screen, and with or without the tangible aspect) will further improve the outcome of a user test should be investigated further.

The PIPC method is a good method to be used by practitioners because the pictures help to explain the different types of problems that children can experience more clearly than verbal
explanations alone. Furthermore, the picture cards serve as a memory aid during the test, and children are able to clearly express problems in a non-verbal way. Therefore, the number of explicitly indicated problems is higher than with standard thinking-aloud.
Testing with children in practice

This chapter presents an extension and revision of some of the existing guidelines proposed by Hanna et al. (1997) for practitioners of how to prepare and conduct user tests (of computer games) with children between five and seven years old (over four and under eight). The seven questions forming the framework for user tests will be used to structure these guidelines: 1. why are we testing, 2. who will be the test participants, 3. where will we test, 4. what will we ask the test participants to do, and 5. how will we organise the test session(s), 6. the data analysis, and 7. the reporting. The recommendations and guidelines are based on the results of the experiments described in this thesis and the experiences of performing usability tests with many children of this age group, testing different adventure games in the usability lab and at schools. New issues that are discussed are the number of test participants, the selection of effective test participants, the use of tasks, and the behaviour of the test facilitator towards the children during the test. Furthermore, a new method is proposed, called the Problem Identification Picture Cards method, to achieve a high level of children’s input during the test.

The following publication is a shorter version of this chapter: (Barendregt and Bekker, 2005d)
7.1 Guidelines to conduct usability tests with children

Although many books have been written on how to conduct user tests with adults (Rubin, 1994, Nielsen, 1993) it is only recently that attention has also been given to user testing with child participants. Hanna, Risden, and Alexander (1997) created a document with guidelines for usability testing with children and this is still one of the very few and most frequently cited papers on this subject. However, their advice is based on experience, not on experimental studies like the ones performed in this thesis. In this chapter some new guidelines for usability testing of computer games with children are given, based on the experiments described in the previous chapters and the personal experiences in the lab and at schools. Furthermore, some guidelines are slightly revised or elaborated for the specific situation of testing computer games with children between five and seven years old.

The guidelines in this chapter are organised according to six of the seven questions in the framework for user testing discussed in Chapter 1: why are we testing, who will be the test participants, where will we test, what will we ask the test participants to do, and how will we organise the test session(s), the data analysis, and the reporting. For the question why are we testing, no guidelines are given by Hanna et al. (1997) and the research in this thesis also gave no rise to any guidelines. For the question of how we will organise the reporting there are no guidelines from Hanna et al. (1997), and because in this thesis the results were never officially reported to the developers of the game there is also no additional advice. For the other questions the advice of Hanna et al. (1997) will be summarised first. The additional or extended advice, based on the experiments in this thesis and the personal experiences from conducting the user test for this thesis, will be presented after each summary.

7.2 Who will be the test participants?

7.2.1 Existing guidelines

1. Screen children for testing that have at least some experience with the computer.
2. Do not include children in testing who have too much experience with the computer.
3. Do not use your colleagues’ children as usability participants.
7.2.2 The number of children

One of the first decisions when planning a user test with children is how many participants to employ in order to detect a sufficient percentage of problems. The formula $1 - (1 - p)^n$, where $n$ is the number of test participants and $p$ is the average detection rate of a given problem, is often used to calculate how many participants are needed to detect a pre-set percentage of all problems (Nielsen, 1994), for example 80%, 90% or 95%. When using only 5 participants, which is sometimes given as a rule of thumb (Nielsen, 1994), an average detection rate of 0.42 is necessary to find about 80% of the problems.

However, detection rates are often much lower (Lewis, 1994, Bekker et al., 2004) and in such cases many more test participants are needed to uncover for example 80% of all problems than the five recommended by Nielsen. In our experience the same is true for children as participants in user tests of computer games. The average detection rate ranged from 0.12 to 0.14 in the experiments in this thesis, which means that eleven to thirteen children would be needed as test participants in order to detect 80% of the problems. To detect 90% of the problems 16 or 17 children would be needed.

---

Note however, that it is not possible to decide exactly how many problems there are; each additional participant can experience new problems that were not experienced by any of the other participants. The formula $1-(1-p)^n$ is an asymptotic formula that approaches 100% with an infinite number of participants.
Even when dismissing all problems detected by only a single child (associated with a low probability of detection), the average detection rate was only 0.19. Altogether, practitioners should be aware of the possibly low average detection rate when testing computer games with children and use target percentages in combination with their budget limitations to select an appropriate number of test participants.

### 7.2.3 Selecting children

Especially when the budget for testing is limited and, subsequently, not many children can be included in the user test, it is crucial to get as much information out of each child as possible. In order to increase the amount of information that an evaluation yields about problems that need to be fixed it is recommended to include only those children who will ostensibly experience many problems, as well as being able to verbalise them adequately.

The research reported here showed that scores on a small set of personality characteristics can be used to predict which children will experience many problems, and, which children will verbalise many problems. In personality theory five main personality characteristics are commonly distinguishable. In this research the Blikvanger questionnaire (Elphick et al., 2002) for parents was used. This questionnaire assesses the five main personality characteristics by dividing them into eighteen sub characteristics. Scoring high on ‘Curiosity’ which is a sub characteristic of the main characteristic ‘Intelligence’ is a good indicator of the
number of problems encountered. Scoring high on the main personality characteristic ‘Extraversion’ in combination with a low score on ‘Friendliness’ is a good indicator of the percentage of verbalised problems.

The Blikvanger is a questionnaire which may only be employed by registered psychologists or under supervision of a registered psychologist. Although there do exist other questionnaires to assess the relevant personality characteristics, it is not yet investigated whether they have the same predictive quality as the Blikvanger.

7.2.4 Children with disabilities

Sometimes there may be children in a school class that have serious learning disabilities or physical problems that prevent them from participating in the same way as other children. For example, in the class that participated in the experiment in Chapter 5 there was a girl who suffered from Down’s syndrome. Her parents said that she liked to participate but that they could imagine that this was a problem. Based on the advice of the teacher and the parents, this girl participated in a test session just like all the other children in order to make her feel part of the group. However, her data were not used in the analysis. Although it costs some extra time to perform test sessions with children of whom the data will not be used it is a good way to prevent discrimination of disabled children. Furthermore, most of the time is spent on the analysis of the test sessions not on conducting them; therefore it is only a relatively small effort to conduct some additional sessions.

7.3 Where will we test?

7.3.1 Existing guidelines

1. Make the lab a little more child friendly.
2. Use laboratory equipment as affectively yet unobtrusively as possible. Place microphones close to the children and opt for smaller ones over larger ones.

7.3.2 Testing at schools

The guidelines of Hanna et al. (1997) focus only on usability tests performed in the lab but most of the experiments in this thesis were conducted at schools. This has the advantage that parents do not have to bring their children to the lab and stay for the duration of the test or pick their children up after the test. More parents will therefore be willing to let their children participate, and it is easier to plan the sessions over the day. A disadvantage is that the
teachers (and the parents) have to be convinced of the acceptability of having children miss classes. To conduct the experiments for this thesis it appeared to be easier to get teachers to accept this for children of five years old (usually group two in the Dutch school system) than for children of six years old (usually group three in the Dutch school system).

7.4 What will we ask the test participants to do?

7.4.1 Existing guidelines

1. When planning series of tasks, switch the order around for different children so that the same tasks do not always come at the end of the test when children are tired.
2. Children older than four or five can be expected to perform specific tasks in addition to free exploration of the product. However, it will be necessary to break down the tasks into smaller segments than for adults.
3. Do not ask children if they want to play the game or do a task, because this gives them the chance to say no.

7.4.2 Tasks or free play

The experiment in Chapter 2 showed that to detect usability and fun problems in a computer game in a realistic situation, it is necessary that children are allowed to play the game freely for at least part of the test session, as opposed to working on tasks. For specific functional parts it can be useful to add some small tasks, e.g. to test whether children know how to turn the volume down. However, a risk of giving tasks is that they can give away information about the game, which the children otherwise may not have found. For example, in ‘Milo and the Magical Stones’ (MediaMix, 2002a) there is a map to navigate more easily from one part of the game to another without having to repeat already finished games. Most children did not notice this functionality although it was explained in the introduction, resulting in frustration about having to repeat games to go to previously visited screens. When one of the tasks would have been to find the map and use it, the children would not have shown this frustration afterwards. This task should therefore only be given after the period of free play.

7.4.3 Making the children stop

Usually each test session has a fixed duration. When the time is up the child has to stop playing the game, but especially with games that children find engrossing it is sometimes hard to make them stop, as was experienced in some of the pilot studies. When the session is
almost over make sure the facilitator warns the child that he/she has to stop in a couple of minutes. Be firm and say something like: “We are going to stop in several minutes.” When you have some time to prolong the session and the child has almost reached a certain sub (goal) at the moment he/she has to stop, it can be helpful to say that the child can play until this (sub) goal is reached. This way the child will experience the end of the test session as less forced and therefore more pleasant, which can make him/her more inclined to participate in a user test again.

7.5 How will we organise the test session(s)?

7.5.1 Existing guidelines

1. If possible, find out what input device children use at home and set it up for them before the test.
2. Schedule the lab for an hour per child. Children up to five years old will be able to concentrate for about 30 minutes but need time to look around and play. Give yourself plenty of time between children and do not plan too many children per day.
3. Establish a relationship with children when you first meet them by engaging them in some small talk to find out more about one another.
4. Explain confidentiality agreements by telling that designs are ‘top-secret’.
5. Have a script for introducing children (and the parents when present) to the testing situation.
6. Set children’s expectations appropriately for what they will be doing during the usability session. Many children will expect to see a piece of finished software and may be disappointed when they are presented with a paper prototype. Explain to them why it is important to get their feedback at the current stage of development.
7. Show children and parents around the lab, including behind one-way mirrors.
8. Younger children (up to seven- or eight-year olds) will need to have a tester in the room with them.
9. Younger or shyer children may be uncomfortable being alone with the tester (Hanna et al. let children under five always have their parents with them throughout testing).
10. If siblings accompany children to a test, make sure they stay in the observation area or another separate room for the duration.
11. It is a good idea to prepare a script of hints that offer varying levels of support for performing given tasks.

12. Children are often used to working or playing with new computer products with others. They are used to asking help from others. You will need to redirect their questions by questions of your own.

13. If children begin to stray from the computer, they should gently be reminded to pay attention to the computer.

14. A way to encourage young children to try an activity they are not immediately attracted to is to pretend that you need help doing it.

15. If the test will run longer than 45 minutes, they should be asked if they want to take a short break at some point.

16. For children who are struggling to read words and numbers you may have to read items for them.

17. Keep children feeling encouraged by offering generic positive feedback, in case they are feeling they are failing at figuring out the software.

18. Younger children’s responses to questions about whether they liked something or not may not be very reliable, because these children are eager to please adults. Older children may be able to give reliable ratings about aspects of the software. It can be helpful to use a vertical scale with a smile face on top and a frown face on the bottom end to make the end markers clear.

19. After testing, reward children by commenting how helpful they were. Explain to them that all their hard work helped you to see exactly which things needed to be fixed.

20. For gratuity, parents and children often appreciate a choice of a gift certificate to a local toy-store or movie theatre in addition to standard software or payment options.

7.5.2 Allowing children to practice before the actual test

The numbers and types of problems and the severity or the detected problems will depend on the amount of practice children have with the tested game before they participate in the test. As described in Chapter 4 knowledge and judgment problems occur most often during first use. These problems are related to knowing what to do in the game and understanding its feedback. Control problems occur more often when children have practiced with a game. These problems are related to explanations and feedback that take too much time and cannot
be interrupted. Furthermore, problems caused by too high a challenge level also occur more often during first use.

Because many knowledge problems are quite serious and require assistance to the child to overcome them, it is more important to test during first use than after the child has practiced with the game. Control problems, which are experienced more often when children have practiced with a game, are merely annoyances. Therefore, testing after children have gained some experience with the game to detect control problems does probably not justify the costs. Challenge is something that can make or break a game; there appears to be a very delicate balance between the right challenge level and either a too high or too low challenge level. To determine whether the challenge level of specific parts of the game is appropriate it is advisable to give children some opportunity to practice and to test those parts again after practice.

7.5.3 Using a script to explain the purpose of the test

Although it is useful to have a script to introduce the children and their parents to the test situation as advised by Hanna et al. (1997) in guideline 5 in section 7.5.1, it is not advisable to refer too strictly to this script. It should be something that the facilitator uses in a very natural way. During the first few sessions of a series of tests I often tried to adhere strictly to my script but it always seemed that the children could pick up my nervousness and unnaturalness and would become very quiet, which is something you don’t want. Only after several tests would I get comfortable and less concerned about whether I had said everything in the right order, which almost always had a positive effect on how open the children were towards me.

7.5.4 Using a script during the test

Testing with young children is very unpredictable. Sometimes children will start to cry, or talk about their health or ask questions about other children. The facilitator should be prepared for this and should be able to improvise. Therefore, strictly following a script for when to say what to a child will often not be possible and is not advisable. For example, in most tests for our research the script was that children would not get help if they did not ask for it repeatedly. However, one girl in our pilot test first stared at the screen for a long time, when I finally asked her whether she knew what to do she started crying and indicated that she had no idea what to do. To make her feel more comfortable I told her that we would play
the game together and I took over the mouse. When she started to feel more comfortable again and was telling me what we should do I gave her back the mouse and she continued playing the game on her own. In this case I did not stick to the script but I think the remainder of the test was still very valuable.

7.5.5 Giving help

Children can sometimes become really sad when they do not understand what has happened or how to proceed. It is therefore very tempting for a facilitator to give help to children when they get stuck and ask for help. Hanna et al. (1997) advise in guideline 12 in section 7.5.1 to redirect questions from children but in guideline 11 they advise to make a script of hints. As argued in Chapter 2, giving help may also be beneficial for the number of problems one can discover. Altogether it is almost inescapable that the facilitator will give help when children ask for it. However, the facilitator should first encourage the children to try a bit longer. This will make it easier to determine the severity of a problem. Furthermore, the facilitator should make sure that help is given at the right level, as explained in section 2.5.1. These levels relate to the phases in the Interaction Cycle which describe how users interact with computer games:

- Determining the goal
- Translating the goal into actions
- Performing the physical actions
- Understanding the feedback
- Evaluating the outcome
- Assessing the motivation to continue playing the game

When the facilitator is not sure what the right level of help is, he/she should start by asking whether the child understands the goal, whether the required actions are clear, and so on. This holds especially for problems that seem at a first glance to be related to the later phases of the Interaction Cycle, like understanding the feedback and evaluating the outcome. Here the facilitator should certainly make sure that the goal and the translation of the goal into actions are understood before giving help about the feedback or the outcome. The advice is thus to first redirect questions from children by asking other questions (as advised in guideline 12 in section 7.5.1) and then give answers to the question at the appropriate level.
7.5.6 Thinking aloud

The thinking aloud technique is often used with adults. During the test the participant is asked to verbalise his/her thoughts while using the product. Young children are often not very good in thinking aloud. One of the reasons is that it is unnatural to talk to no-one in particular. During one of our tests one child when asked to think aloud responded with: “But to whom should I be talking, to you?” This is a clear example of Boren and Ramey’s (2000) position that people always need a conversational partner to be able to think aloud. The facilitator should therefore respond naturally to remarks of the children without biasing them. For example, children will often be very enthusiastic when reaching a sub goal, and they will say things like “I did this very well, didn’t I?”. Usually, I respond to these remarks but I try to keep it short in order to keep the child playing, e.g. “Hm hm, very well”. Although co-discovery is another technique to give children a conversational partner, it should be noted that young children often not really co-operate with each other (van Kesteren et al., 2003) and that pairs may tend to discuss topics unrelated to the product under evaluation (Als et al., 2005).

Furthermore, prompting children to keep talking when they keep silent is often not very useful. Sometimes children will respond to this request but it is very doubtful whether their response is valid because it seems that they are just making something up to say. After a single response most children remain quiet again.

7.5.7 Using the Problem Identification Picture Cards method

The aim of a user test with children is to detect as many problems as possible in the tested product with as much explanation as possible from the children. The Problem Identification Picture Cards Method described in Chapter 6 was proposed to support children express more problems explicitly during the user test. The PIPC method makes use of a box with picture cards symbolizing different types of problems that children can encounter while playing a game. These picture cards were used to explain the purpose of the test and served as a reminder during the test. Finally, the picture cards could be used by less articulate children to express problems clearly in a non-verbal way by putting a picture card in the box. An experiment showed that children indicated more problems explicitly with the picture cards than without the picture cards, without decreasing the number of verbalised problems. In addition, the children liked to use the picture cards.
In practice I would certainly recommend using pictures to explain what information you would like to get from the children because it is much easier to keep children's attention during the explanation. The facilitator can engage the children more during the introduction and it is a good opportunity to establish rapport. Furthermore, I would also strongly recommend keeping the pictures within children’s focus of attention during the test to serve as a reminder of the concepts. However, making the children put the picture cards in a box is probably not very beneficial because it takes too much time to pick up a picture card and put it in the box, and is therefore too distracting. Other less distractive means to select pictures should be provided.

7.5.8 Questionnaires
Often it may be necessary to ask children some questions after the test session, for example what they think of the game as a whole or how much they liked to participate in the test. Unfortunately, asking children these additional questions after the session is difficult. For the experiment in chapter 6 children were asked only the two questions from Appendix 2, but many children were not really interested anymore and wanted to leave once they had to stop playing the game. The questions should therefore be short and easy, and there should not be many questions.

7.5.9 Gifts
In some studies the children are given a small gift after the test session to show appreciation for their cooperation and make them willing to participate again. In that case, make sure you have some extra gifts ready. Often parents will bring siblings or friends to the lab because they want to wait for their child during the session and can not leave the other child(ren) unattended. It would be disappointing if the other children would not get a gift, especially because they are sometimes already disappointed that they cannot participate (e.g. because they are too young).
If you want to give the children something to eat, get some alternatives for those children who cannot eat wheat, milk, chocolate, sugar, peanuts etc. due to allergies.
When testing at a school instead of a usability lab (which is the default situation in Hanna et. al (1997)) it is better to give a present to the whole group rather than to the individual children only. This way, the children who were not able or not allowed to participate by their
parents will feel less disappointed. Teachers will be more inclined to cooperate in future tests when they know that the experience will be pleasant for the class as a whole.

7.6 How will we organise the data analysis?

7.6.1 Existing guidelines

There is only one existing guideline about how to perform the data analysis of user tests with children from Hanna et al. (1997):

1. Gauge how much children like a program by observing signs of engagement such as smiles and laughs, and signs of disengagement such as frowns, sighs, yawns, or turning away from the computer.

7.6.2 Using a coding scheme

The guideline given in the previous section is rather vague and does not provide much guidance for how evaluators should set up the data analysis. The results of the analysis of the recorded video material should be reliable over time and over different evaluators. It is therefore important to describe the types of behaviour, actions, and verbalisations that evaluators should notice. In Chapter 3 of this thesis an extensive coding scheme was described and evaluated. The purpose of this coding scheme was to detect both fun and usability problems. We have found this coding scheme very useful for the experiments in this thesis. Furthermore, it was easy to teach the coding scheme to others. We advise practitioners to use a coding scheme similar to ours, probably with some adaptations to suit the particular aims of the test. For example, when analysing video data of children playing other types of games it may appear that some breakdown indication types are missing. These breakdown indication types could be added to the coding scheme. However, the reliability of the adapted coding scheme should be retested as was done in section 3.4. It requires some training before evaluators use such a coding scheme in a reliable way, so make sure the coding scheme is used to analyse a number of test tapes before the actual analysis is performed. The evaluators should discuss their coding of the test tapes in order to clear up misunderstandings about what each breakdown indication type means, and when a breakdown indication should be coded.
7.6.3 Number of evaluators

As Jacobsen et al. (2003) already described, no single evaluator will detect all problems when analysing a videotaped usability test session. This also holds for the analysis of user tests with children, and this is the reason that for most of our research the analysis was done by two evaluators. However, there is another reason to include more than one evaluator. Young children are often not very successful in thinking aloud. Therefore, the evaluator has to interpret a lot of non-verbal behaviour and unfinished or vague sentences. When doing this alone the evaluator is often not able to determine the exact problem or alternative views. Discussing certain behaviours and verbalisations with another evaluator to create clearer problem reports, is definitely worthwhile (Jordan and Anderson, 1995).

7.7 Discussion and conclusions

The guidelines and advice presented in this chapter are based on the experiments performed in four experiments with about 80 children between five and seven years old, participating in the evaluation of computer games. This is one of the youngest age groups to perform observational evaluations with. Only a minor part of these guidelines is exclusively relevant for children in this age group, most guidelines can easily be applied to testing with older children and sometimes even with adults. For example, using pictures to explain the topics of interest for the test is probably also beneficial for older children because the combination of auditory and visual information adheres to the principles of multiple resources and redundancy gain (Wickens et al., 2004) and therefore facilitates comprehension. Some other advice especially aimed at this age group is about using very short questionnaires. The older children get, the more they are used to getting tasks, therefore they will probably be more inclined to obey to the demands of the facilitator. Furthermore, the attention span of children increases with age, making them able to stay focused for longer periods.

The advice about not giving tasks is clearly related to evaluating computer games. For other types of products it may be necessary to give tasks to make children use the product in a representative way. Furthermore, the pictures used in the PIPC method are clearly aimed at finding problems in computer games; for other products there should probably be different picture cards representing the types of problems that can occur in such products.

In this chapter the practical advice based on all the experiments presented in this thesis is gathered to help practitioners use the results in a pragmatic way. Although there may be many more issues that could be added to the guidelines presented here, we think that
especially the advice based on the experiments forms a valuable extension to the guidelines of Hanna et al. (1997).
Discussion and conclusions

In this final chapter the main conclusions of the experiments described in the previous chapters are presented. Subsequently, attention is given to a possible limitation of this thesis. Finally, several directions for further research are proposed according to the seven questions forming the framework for user tests used throughout this thesis.
8 Discussion and conclusions

8.1 Recapitulation

The main research question of this thesis was: How can we increase the amount of relevant information yielded by observational evaluations with children between five and seven years old, about usability and fun problems that should be fixed in order to improve games? In order to answer the main research question four experiments were performed with the following underlying research questions:

1. Should children participants perform tasks during the evaluation? (Chapter 2)
2. Should children participants have practiced with the game before the actual test is conducted in order to find the problems that need to be fixed? (Chapter 4)
3. Is it possible to select children who will make good participants based on their personality characteristics? (Chapter 5)
4. Is there an effective alternative method for the think-aloud method for evaluating games with children? (Chapter 6)

In order to answer these questions a reliable coding scheme to detect usability and fun problems in video data of children playing games was needed. The development and evaluation of a coding scheme to detect behaviours that signal problems was described in Chapter 3. This coding scheme can be used as a basis for other researchers and practitioners wishing to analyse video data of children playing adventure games. Furthermore, consideration was given in Chapter 2 to the problem of prompting children to thinking aloud and providing the right kind of help during the evaluation. These considerations have not yet led to research questions or clear answers but they could give inspiration to other researchers wishing to investigate the evaluation of games with children. Some practical recommendations for practitioners wishing to evaluate computer games with children were given in Chapter 7.

In this final chapter the conclusions to the four underlying research questions are presented once more in Table 8-1. In this table first the main research question of each experiment is presented. Subsequently, the sub questions that helped to answer each main question are presented. A limitation to the presented work is given in section 0. The chapter is concluded by looking at several further research areas in section 8.3.
## Discussion and conclusions

Table 8-1 Main research questions and sub questions followed by the answers presented in this thesis.

<table>
<thead>
<tr>
<th>No</th>
<th>Question</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Should children participants perform tasks during the evaluation of a computer game?</td>
<td>No, the goals set by the tasks make children try harder to finish each task than they would normally have done when playing the game. Therefore, it influences children's natural playing behaviour.</td>
</tr>
<tr>
<td>1a</td>
<td>Is there a difference in the number of verbal and non-verbal indications between children performing a task and children playing freely?</td>
<td>No, there was no significant difference.</td>
</tr>
<tr>
<td>1b</td>
<td>Is there a difference in the number of detected problems between children performing a task and children playing freely?</td>
<td>No, there was no significant difference.</td>
</tr>
<tr>
<td>1c</td>
<td>Is there a difference in the number of visited screens between children performing a task and children playing freely?</td>
<td>Yes, children having to perform tasks visited significantly more screens than children playing freely.</td>
</tr>
<tr>
<td>1d</td>
<td>Is there a difference in the number times asked for help between children performing a task and children playing freely?</td>
<td>Yes, children having to perform tasks asked significantly more often for help than children playing freely.</td>
</tr>
<tr>
<td>2</td>
<td>Should children participants have practiced with the game before the actual test is conducted in order to find the problems that need to be fixed?</td>
<td>No, especially knowledge problems occur often during first use and can not be solved by the children themselves; they need help from somebody else. Therefore, it is necessary to evaluate first use.</td>
</tr>
</tbody>
</table>
## Discussion and conclusions

<table>
<thead>
<tr>
<th>No</th>
<th>Question</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a</td>
<td>What kind of problems do children encounter when they play a game for the first time during an evaluation and what kind of problems do children encounter when they have practiced with the game?</td>
<td>The first time that children play a game they will have more knowledge problems, judgement problems, and problems related to a too high challenge level than after some practice. After children have practiced with a game they will experience more control problems than when they played the game for the first time.</td>
</tr>
<tr>
<td>2b</td>
<td>Does the severity of problems that children encounter both during first time and after some practice change?</td>
<td>The impact severity of problems that are experienced during first use and after some practice decreases. The frequency severity of problems remains the same.</td>
</tr>
<tr>
<td>2c</td>
<td>Do the standard usability metrics Efficiency, Effectiveness, and Satisfaction change when children practice with a game?</td>
<td>Yes, the usability metrics Efficiency, Effectiveness, and Satisfaction all increase when children practice with a game.</td>
</tr>
<tr>
<td>3</td>
<td>Is it possible to select children who will make good participants based on their personality characteristics?</td>
<td>Yes, Curiosity, Extroversion, and Friendliness are good predictors for the effectiveness of children as participants in an evaluation. Selecting these children will result in detecting representative problems for the whole user group.</td>
</tr>
<tr>
<td>3a</td>
<td>Is it possible to predict which children will encounter many problems?</td>
<td>Yes, children who score high on Curiosity according to the Blikvanger questionnaire encounter the highest number of problems.</td>
</tr>
<tr>
<td>3b</td>
<td>Is it possible to predict which children will verbalise many problems when they encounter them?</td>
<td>Yes, children who score high on Extroversion and low on Friendliness according to the Blikvanger questionnaire verbalise the highest percentage of problems that they encounter.</td>
</tr>
</tbody>
</table>
8 Discussion and conclusions

<table>
<thead>
<tr>
<th>No</th>
<th>Question</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>3c</td>
<td>Are the problems detected by these most-promising children representative for the problems detected by all children?</td>
<td>Yes, based on the formula $1 - (1-p)^n$ it is possible to predict the percentage of detected problems for a smaller subgroup of the group of participants. The subgroup of most promising children finds expected percentages of problems of high and medium severity, and only a slightly lower percentage of the low severity problems than expected.</td>
</tr>
<tr>
<td>4</td>
<td><em>Is there an effective alternative method for the think-aloud method for evaluating games with children?</em></td>
<td><em>Yes, the Problem Identification Picture Cards method is a promising method to make children express more problems, and children like to use this method during an evaluation.</em></td>
</tr>
<tr>
<td>4a</td>
<td>Is it possible to make children express more problems explicitly by means of the Problem Identification Picture Cards method?</td>
<td>Yes, when children use the Problem Identification Picture Cards method they express more problems explicitly (verbally, with only a Picture Card, or with a combination of a verbalisation and a Picture Card) than when using thinking-aloud.</td>
</tr>
<tr>
<td>4b</td>
<td>Do children still verbalise as many problems with the Problem Identification Picture Cards method as without this method?</td>
<td>Yes, there is no significant difference between the numbers of verbalised problems with and without the Problem Identification Picture Cards method.</td>
</tr>
<tr>
<td>4c</td>
<td>Do children like to use the Problem Identification Picture Cards method?</td>
<td>Yes, there are more children who like to perform another evaluation with this method than without this method.</td>
</tr>
</tbody>
</table>

In summary, this thesis has provided insights in three important factors to increase the amount of relevant information yielded by observational evaluations with children between five and seven years old, about usability and fun problems that should be fixed in order to improve games: the selection of children based on personality characteristics, asking children
to practice with the game before the test session, and a novel way to make children express more problems during the evaluation. These insights were obtained by performing three rigorous experiments, which is not yet very common in the fields of child-computer interaction and games. Based on the results of this thesis future evaluations of computer games with children can be made more effective in terms of the numbers of detected problems and the information obtained from the children about these problems.

### 8.2 Limitation of the research: problem fixing and redesign

The main result of an evaluation of a computer game is a list and a description of the problems that need to be fixed by the developers. Fixing these problems for the purpose of improving the quality of the game is the final aim of the evaluation. The focus of this thesis was on how to detect problems and increase the amount of information obtained from the children about these problems. The focus was not on how these problems should eventually be fixed by the developers once they were detected. When determining the usefulness of an evaluation method it is also important how well the method helps to redesign the product. The assumption regarding redesign and problem fixing was that information expressed by the children themselves is more convincing and has a higher explanatory value to the developers who have to improve the product than information obtained from only observing nonverbal behaviour of children. However, this may still not be enough to redesign and improve the product in a later stage. Therefore it is not guaranteed that the results of this thesis e.g. selecting most-promising children or using the PIPC method are enough to deliver benefit to the developers of a game. However, the results of Chapter 4 indicate that fixing problems has a positive influence on efficiency, effectiveness, and satisfaction. Although in this experiment no problems were fixed through redesign of the game, the help provided by the researchers during the first test session solved many problems. Therefore, it is expected that fixing the detected problems in the software will also increase the quality of the game.

Evaluators may have to talk with the children about the problems they encounter to be able to diagnose a problem well enough to advice the developers. This is certainly a long way from Ericsson and Simon’s (1984) theory of gathering data through verbal protocols, and the chance that the evaluator biases the participants, invalidating the results, seems very high. Another approach is to fix problems on the spot and re-evaluate the new version of the game with another participant. Such a method is the RITE method which is described by Medlock
Discussion and conclusions

et al. (2002). Future research is needed to determine the possible benefits and risks of these approaches.

8.3 Future work

The framework for conducting user tests will be used once more to give some directions for future work based on the results or discussions in this thesis. However, before using this framework again it should be noted that one additional question could be added: what are we testing? The research in this thesis was aimed at increasing the information about usability and fun problems in computer games of the adventure genre. However, some other genres do exist for young children, for example simulation, racing games, and action games. It would be interesting to investigate whether the proposed coding scheme works for these other games genres, or whether it may be necessary to include additional concepts based on observations of children playing these games and remove others. The same holds for the symbols used on the picture cards (they are of course closely related to the coding scheme) in the PIPC method. The influence of practice on experienced problems and their severity in games of other genres than adventure games is also worth investigating.

8.3.1 Why are we testing?

In this thesis the main aim was to detect relevant usability and fun problems and obtain information about these problems from the children. As stated in Chapter 1, the focus is on observational evaluations. However, other types of evaluations do exist: analytic and inquiry methods. In chapter 7 a guideline was given about using very short questionnaires for young children. This guideline was based on experience not on the results of an experiment. It would be useful to investigate how questionnaires for young children should be made more interesting. The Berkeley Puppet Interview method to assess children’s self-perception by Measelle et al. (1998) is an example of a way to make a questionnaire better suitable for children because it reduces the reluctance of children to give negative answers. This could be a good starting point to develop questionnaires about computer games.

8.3.2 Who will be the test participants?

Many of the issues discussed in this thesis are not exclusively applicable to children or computer games and could be used as a basis to investigate user testing with older participants, e.g. personality characteristics of adults may also play an important role in how
well someone will be able to participate in a user test and the behaviour of adults with a computer game will probably also change when they have practiced with a game. Even the use of pictures to remind adults of the topics of interest for the evaluation may be beneficial. For older children and adults the concepts used for the picture cards can probably be more fine-grained so that it is possible to distinguish different type of problems more easily.

8.3.3 Where will we test?

In this research the user tests were all conducted at the lab or at schools. Some pilot tests were conducted at the homes of the children, but it has not been investigated what the influence is on the results of the evaluation. Markopoulos et al. (2005) have written about an attempt to include parents as evaluators for children playing a game in the home environment, but further research could be useful.

8.3.4 What will we ask the test participants to do?

In this thesis the children were allowed to play freely. However, it may sometimes be necessary to test specific parts of a game by asking the children to perform some tasks in combination with free exploration. Hanna et al. (1997) already hinted that tasks for young children should be broken down into smaller segments than for adults, but it is not clear what the size of the segments should be for each age group.

8.3.5 How will we organise the test session(s)?

Other evaluation methods could be proposed to increase the effectiveness of an evaluation e.g. is it effective to involve parents as facilitators or evaluators (Markopoulos et al., 2005)? Furthermore, it would be very interesting to see what happens to problems of different types over a longer time of practice than in the experiment described in Chapter 4. What kind of problems will remain and which problems make children dislike a game eventually? Possibly some guidelines for game designers to create games that remain fun to play over a longer time can be drawn up from this further research. Another logical step would be to improve the PIPC method, as was mentioned before. It would certainly be interesting to investigate alternative ways to use the symbols as a way to express problems without having to put the cards in a box. For example, by placing the symbols on the same screen as the game and making the children touch them. Another line of research would be to investigate the usefulness of the PIPC method for children with different personality characteristics. Does
the method support children who are good verbalisers in expressing problems even more, or does the method support children who are not very good in verbalising in expressing problems? Regarding the giving of help during the evaluation it would be interesting to investigate the influence of giving help, and giving different kinds of help on how many problems are detected. Furthermore, it is probably necessary to investigate more closely how facilitators should make sure to provide the right kind of help in order not to confuse children, and not give away too much information which could leave some problems undetected. Chapter 2 only provided some first insights on this topic.

8.3.6 How will we organise the data analysis?

In this thesis a coding scheme and a procedure in which evaluators had to discuss their individual results to come to final problem descriptions were developed in order to be able to detect and count problems in a reliable way. The coding scheme was also used by some other students for their work on usability evaluations with children. However, learning how to use a coding scheme takes practice, and finding multiple evaluators willing to analyse the data with such a coding scheme is not always easy. Investigating ways to decrease the learning time and the time to code the test sessions without lowering the reliability would be useful.

8.3.7 How will we organise the reporting?

No research was performed on reporting the results of the evaluations in this thesis, but of course reporting the results to the developers in such a way that they can fix the detected problems and improve the game is very important. One way to communicate problems to the developers is by creating small movies from the test sessions to illustrate each problem. Research on how to create effective problem reports based on the detected problems and the way to communicate them to the developers is very useful.
Bibliography


Barendregt, W., Bekker, M. M., 2005b. Developing a coding scheme for detecting usability and fun problems in computer games for young children. Accepted for publication in Behavior Research Methods, Psychonomic Society Publications.


Hartson, H. R., 2005. Uncovering problems in layers. [Personal communication]


MediaMix, 2002b. Regenboog, de mooiste vis van de zee (Rainbow, the most beautiful fish in the ocean) [Computer software] MediaMix Benelux.


MediaMix, 2004b. Max en het rode fruit (Milo and the red fruit) [Computer software] MediaMix Benelux.


Bibliography


Bibliography


Read, J., 2005. Smileyometer. [Personal communication]


Bibliography


Transposia, 2000. Oscar en de geheimen van het bos (Oscar the balloonist and the secrets of the forest) [Computer software] Lannoo.

Transposia, 2000. Freddi Fish, Dolle Doolhof (Freddi Fish, Silly Maze) [Computer software] Transposia.


Summary

The main research question of this thesis is: How can we increase the amount of relevant information yielded by observational evaluations with children between five and seven years old, about usability and fun problems that should be fixed in order to improve games?

The thesis consists of three main experiments. These three experiments investigate respectively, the influence of personality characteristics of children, the influence of practice with the tested game, and the use of a new evaluation method, called the Problem Identification Picture Cards (PIPC) method. In order to detect fun and usability problems during observations of children playing games, first a coding scheme for problem indicating behaviour was developed and tested. This coding scheme was used as a basis for most of the experiments.

The aim of the first experiment was to determine whether it is possible to predict what types of children would be effective participants during an evaluation. This experiment investigated whether personality characteristics, as measured with a standard personality test for young children called the Blikvanger, can be used to select children who will help to detect many problems, and verbalise these problems. The experiment showed that children who score high on Curiosity will encounter many problems, and children who score high on Extroversion but low on Friendliness will verbalise a high percentage of these problems. An analysis was performed to investigate whether the results obtained with a selection of most-promising children, based on the scores on these favourable characteristics, can be representative for the problems of all children.

The second experiment investigated whether it makes a difference in the outcomes of an evaluation when children experience a game for the first time or when they have practiced with the game for about an hour before the evaluation. Videotapes of children playing a game
Summary

for the first and after they had practised were analyzed and problems were categorised according to two frameworks: the taxonomy of usability problems according to Zapf and colleagues (1992), and a taxonomy of fun problems based on the work of Malone and Lepper (1980, 1987). The experiment showed that for the tested game efficiency, effectiveness, and satisfaction went up when children were more experienced with the game. Furthermore, the number of problems related to a lack of knowledge of the game, not understanding the feedback, and a too high challenge level went down, whereas the number problems related to too long explanations and feedback went up. Finally, the problems that had the highest severity when children played the game for the first time were no longer the problems with the highest severity when children had practiced with the game. Thus, letting children practice with the game before the evaluation takes place can give a different view of the things that need to be changed to improve the game. Because especially knowledge problems can often not be solved by children without getting help from someone else it was concluded that it remains important to test first use.

The last experiment investigated whether the use of a box with picture cards, symbolizing the different types of problems, could increase the number of self-reported problems during an evaluation. The children could put picture cards in the box when they encountered problems. This PIPC method was intended to serve as an addition to the thinking-aloud technique, which children are commonly supposed to use. The experiment showed that with the PIPC method the children verbalised at least as many problems as without the PIPC method. Furthermore, the children would rather like to use the PIPC method for future evaluations than standard thinking-aloud. The main finding was that with the PIPC method the children were able to indicate more problems explicitly than when they had to think aloud without the PIPC method.

All information from these experiments served as input for chapter 7 in this thesis, in which an overview of guidelines to set up an evaluation of a computer game with young children is given. In summary, this thesis has provided insights in three factors to increase the amount of relevant information yielded by observational evaluations with children between five and seven years old, about usability and fun problems that should be fixed in order to improve games: the selection of children based on personality characteristics, asking children to practice with the game before the test session, and a novel method to make children express more problems during the evaluation.
Samenvatting

De centrale onderzoeks vraag voor dit proefschrift was: Hoe kunnen we de hoeveelheid informatie verhogen die wordt verkregen uit observationele evaluaties van een computerspel met kinderen van vijf tot en met zeven jaar oud, over bruikbaarheids- en plezierproblemen die opgelost moeten worden om de kwaliteit van het spel te verbeteren? Het detecteren en beschrijven van bruikbaarheids- en plezierproblemen wordt gezien als een noodzakelijke stap om de kwaliteit van computer spellen voor jonge kinderen te verhogen. De focus van dit proefschrift is de evaluatie van avonturen spellen voor de thuis markt.

Dit proefschrift bestaat uit drie hoofdexperimenten. Deze drie experimenten onderzoeken respectievelijk: de invloed van persoonlijkheidskenmerken van kinderen, de invloed van oefening met het geteste spel, en het gebruik van een nieuwe evaluatie methode, de Problem Identification Picture Cards (PIPC) methode. Om bruikbaarheids- en plezierproblemen te kunnen ontdekken tijdens een observatie is eerst een codeerschema voor probleem aanduidende gedragingen ontworpen en geëvalueerd. Dit codeerschema is gebruikt als de basis voor de meeste experimenten.

Het doel van het eerste experiment was om te bepalen of het mogelijk is om te voorspellen welke types kinderen effectieve deelnemers aan een evaluatie zullen zijn. Het experiment onderzocht of persoonlijkheidskenmerken zoals die gemeten kunnen worden met een standaard persoonlijkheidstest voor jonge kinderen, de Blikvanger, gebruikt kunnen worden om kinderen te selecteren die veel problemen helpen ontdekken en veel informatie over elk ervaren probleem geven. Het experiment toonde aan dat kinderen die hoog scoren op Nieuwsgierigheid veel problemen zullen tegenkomen, en dat kinderen die hoog scoren op Extraversie maar laag op Vriendelijkheid een hoog percentage van de problemen die ze tegenkomen zullen verbaliseren. Een analyse was uitgevoerd om te onderzoeken of de resultaten van een evaluatie verkregen met kinderen van wie op basis van hun
persoonlijkheidskenmerken verwacht kan worden dat ze effectieve deelnemers zullen zijn, representatief kunnen zijn voor de problemen van alle kinderen.

Het tweede experiment onderzocht of het een verschil maakt op de uitkomsten van een evaluatie als kinderen een spel voor de eerste keer spelen of als ze al geofend hebben met een spel voor de evaluatie plaatsvindt. Videobanden van kinderen die een spel de eerste keer speelden en nadat ze met het spel hadden geofend zijn geanalyseerd en de gevonden problemen zijn gecategoriseerd met behulp van twee raamwerken: de taxonomie voor bruikbaarheidsproblemen volgens Zapf en collega’s (1992), en een taxonomie voor plezierproblemen gebaseerd op het werk van Malone and Lepper (1980, 1987). Het experiment toonde aan dat voor het geteste spel de efficiëntie, effectiviteit, en tevredenheid toenamen wanneer kinderen meer ervaring met het spel kregen. Verder namen de aantallen problemen gerelateerd aan een gebrek aan kennis van het spel, het niet begrijpen van feedback, en een te hoog uitdagingniveau af, terwijl de aantallen problemen gerelateerd aan te lange uitleg of feedback juist toenamen. Als laatste waren de problemen die het ernstigst waren als kinderen het spel voor de eerste keer speelden niet meer het ernstigst als ze ermee geofend hadden. Er kon dus geconcludeerd worden dat het laten oefenen met een spel voordat de evaluatie plaatsvindt zeker een andere blik werpt op wat er aan een spel verbeterd moet worden. Omdat vooral kennisproblemen niet goed door kinderen opgelost kunnen worden zonder hulp van iemand anders blijft het toch erg belangrijk om het eerste gebruik te evalueren.

Het laatste experiment onderzocht of het gebruik van een doos met kaartjes waarop plaatjes staan van de verschillende soorten problemen kan helpen om het aantal problemen dat kinderen zelf aangeven tijdens een evaluatie te verhogen. De kinderen konden kaartjes van elk type in de doos doen als ze een probleem tegenkwamen. Deze PIPC methode was bedoeld als een aanvulling op het hardop denken wat vaak gebruikt wordt voor kinderen. Het experiment liet zien dat kinderen met de PIPC methode net zoveel problemen verbaliseerden als zonder de PIPC methode. Verder zouden de meeste kinderen in toekomstige evaluaties liever de PIPC methode gebruiken dan alleen het hardop denken. De belangrijkste bevinding van dit experiment was echter dat kinderen met de PIPC methode meer problemen zelf expliciet aangaven dan als ze alleen hardop moesten denken.

Alle informatie uit de experimenten diende als bron voor hoofdstuk 7 in dit proefschrift, waarin een overzicht van de adviezen om een evaluatie van een computerspel met jonge kinderen wordt gegeven. Samenvattend heeft dit proefschrift inzicht verschaft in drie
Samenvatting

factoren die kunnen bijdragen aan een vermeerdering van de hoeveelheid informatie die verkregen wordt tijdens een evaluatie met kinderen van vijf tot en met zeven jaar oud over bruikbaarheids- en plezierproblemen die opgelost dienen te worden om de kwaliteit van computerspellen te verhogen: de selectie van kinderen gebaseerd op persoonlijkheidskenmerken, kinderen vragen om te oefenen voor de test sessie en een nieuwe methode om kinderen meer problemen te laten aanduiden tijdens de evaluatie.
Curriculum Vitae

10 okt. 1971 Geboren te Purmerend

1984-1985 Voorbereidend Wetenschappelijk Onderwijs (VWO), Elzendaal College, Boxmeer

1985-1988 Voorbereidend Wetenschappelijk Onderwijs (VWO), Thorbecke College Arnhem, in combinatie met Vooropleiding Dansacademie aan de Hogeschool voor de Kunsten, Arnhem

1998-1990 Voorbereidend Wetenschappelijk Onderwijs (VWO), Haags Montessori Lyceum, Den Haag

1990-1995 Studie Informatica, Radboud Universiteit Nijmegen

1995-1996 Functioneel Ontwerper, Rijks Computer Centrum, Apeldoorn

1996-1999 Software/Information Engineer, Kadaster, Apeldoorn

1999-2001 Interaction Designer, Getronics Interedge, Amersfoort

2002-nu Promovendus, Technische Universiteit Eindhoven, Faculteit Industrial Design
Dankwoord

De afgelopen vier jaar heb ik met heel veel plezier gewerkt aan dit proefschrift. Een groot aantal mensen heeft bijgedragen aan het ontstaan van dit proefschrift en aan het plezier dat ik beleefde tijdens het schrijven ervan. Deze mensen wil ik hier graag bedanken.

Als eerste wil ik Tilde Bekker, mijn begeleidster en copromotor, bedanken. Al bij het sollicitatiegesprek leek het me prettig om met je samen te werken, en gelukkig is dat in de praktijk ook gebleken. Jouw opbeurende woorden had ik af en toe hard nodig, ook al dacht je soms misschien dat het weinig zin had ;-) Bovendien heb ik ook inhoudelijk veel steun aan je gehad. Ik hoop dat we kansen vinden om samen te blijven werken na mijn vertrek uit Eindhoven.

Ook wil ik prof. Bouwhuis, mijn eerste promotor, bedanken voor de inspiratie en nuttige discussies tijdens het schrijven van de artikelen en voor het zoeken van uitgebreide informatie over persoonlijkheidstests.

De leescommissieleden prof. Rauterberg, prof. de Ridder en prof. Stappers bedank ik voor het lezen van mijn proefschrift en het aandragen van nuttige adviezen.

Special thanks to prof. Cockton who must have spent weeks on reading my thesis and providing me with so many useful and detailed comments. It has definitely become a much better piece of work thanks to you!
Dankwoord

Ik wil prof. Chris Snijders van de faculteit Technology Management heel erg bedanken voor het steeds weer helder reageren op mijn statistiekvragen.

Mijn kamergenoten Ester Baauw en Wouter Sluis-Thiescheffer wil ik graag bedanken voor de fijne tijd die we samen hebben gehad. Ester, zonder jouw hulp was het niet gelukt! Ik vond onze samenwerking heel plezierig en ik ga je erg missen. Ik ben dan ook erg blij dat je mijn paranimf wilde zijn. Wouter, bedankt voor je steun tijdens het afronden van mijn proefschrift, en bij mijn overpeinzingen om naar Zweden te gaan. Veel succes met het schrijven van jouw proefschrift.

Koen van Turnhout wil ik bedanken voor alle boeiende gesprekken over van alles en nog wat, en voor het aanvoelen van wanneer wat morele ondersteuning nodig was. Je hebt verder ook nog een belangrijke bijdrage aan hoofdstuk 8 geleverd toen ik het op vrijdagavond niet meer zo zag zitten. Ik vind het jammer dat ik je niet echt kan helpen als jij met het afronden van je proefschrift bezig bent, maar bellen of mailen kan natuurlijk altijd!


Verder zijn er de mensen waarmee ik in contact ben gekomen omdat we gedeeltelijk dezelfde onderzoeksinteresses hebben: Arnold Vermeeren, Bieke Zaman, Johanna Höysniemi en Afke Donker. Bedankt voor de nuttige uitwisseling van gedachten en literatuur en het plezierige contact tijdens de conferenties die we samen bezochten.

Anne Jansen wil ik bedanken voor het prettige contact tijdens haar werkzaamheden bij Noldus en voor de morele ondersteuning bij de beslissing om naar Zweden te gaan na mijn promotie.
Dankwoord

Bij het ontwikkelen van de PIPC methode heb ik erg nuttig advies van speltherapeute Lenette Raaijmaker ontvangen. Heel erg bedankt dat je zomaar voor een onbekende onderzoeker tijd vrij wilde maken.

Mijn ouders wil ik bedanken voor het heilige vertrouwen dat ze altijd in mij hebben gehad, en vooral voor de lieve telefoontjes in de laatste weken voordat mijn proefschrift geaccepteerd werd. Ik vond dat erg fijn en bijzonder.

Als laatste bedank ik Halbe Huitema. Hoewel je zelf vindt dat je niet veel hebt bijgedragen heb je er wel altijd voor gezorgd dat ik een liefdevol thuis had, en daar wil ik je toch zeker voor bedanken!
Appendix 1 Previous experience questionnaire and permission form

In Chapters 4 and 5 the parents of the children involved in the experiment had to complete a short questionnaire about previous experience of their child with computers. Furthermore, the parents had to give a written permission for their child to participate in the experiment. The Dutch questionnaire and permission form are given below.

Naam kind: ...............................................................................................................................  
Geboortedatum: .........................................................................................................................  
Naam juffrouw: ............................................................................................................................  

1. **Heeft uw kind naar uw weten al eerder het spel ‘Max en de Toverstenen’ gespeeld?**
   - ja *
   - nee

* Wanneer u deze vraag met ja heeft beantwoord kan uw kind helaas niet deelnemen aan ons onderzoek. U hoeft het formulier en de vragenlijst niet verder in te vullen.

2. **Hoe vaak maakt uw kind gemiddeld buiten school gebruik van de computer?**
   - Iedere dag

---

8 In this questionnaire the gender of the child was not asked because this was expected to be clear from the gathered video material.
Appendix 1

- 4-6 dagen per week
- 1-3 dagen per week
- 1-3 keer per maand
- Minder dan eens per maand

3. Wanneer uw kind gebruik maakt van de computer, hoe lang duurt dit dan ongeveer?

- Minder dan 15 minuten
- 15 tot 30 minuten
- 30 minuten tot 1 uur
- 1 tot 2 uur
- Meer dan 2 uur

Hierbij verleen ik mijn toestemming om mijn kind deel te laten nemen aan dit onderzoek.

Naam: ..................................................................................................................
Datum: ..................................................................................................................
Plaats: ...................................................................................................................

.............................................

Handtekening
Appendix 2 Preference for game and method questionnaire

In Chapter 6 the children involved in the experiment had to indicate which game they preferred and which method they preferred for performing future evaluations with other games. To make sure each game and each method was mentioned equally often as the first choice and as the second choice, two different orders were used in the answers to these questions. One of the two questionnaires is given below; the other has the answers to the questions in the reverse order. The facilitator read out the questions and the child could answer with a cross in one of the boxes.
Naam:

1. Welk spel vind je het leukst?

Max

![Max cover](image)

IJsbeer

![IJsbeer cover](image)

2. Als je nog een keer een ander spelletje zou mogen testen, zou je het dan het liefste met of zonder de kaartjes doen?

Zonder de kaartjes

Met de kaartjes

![Zonder de kaartjes](image)

![Met de kaartjes](image)
Appendix 3 Consent form

In all experiment the parents had to give their consent and indicate how the video data of their child could be used. The Dutch consent form is given below.

Wilt u een van onderstaande opties aankruisen?

☐ Ik ga er mee akkoord dat de opnames (geluid en beeld) die van mijn kind gemaakt worden tijdens deze test alleen intern gebruikt worden (communicatie naar andere betrokkenen bij dit promotie onderzoek)

☐ Ik ga er mee akkoord dat de opnames (geluid en beeld) die van mijn kind gemaakt worden tijdens deze test zowel intern als extern gebruikt kunnen worden (communicatie naar derden, bijvoorbeeld in presentaties)

Datum:

Naam kind:

Naam: Handtekening: