Predicting effectiveness of children participants in user testing based on personality characteristics

Wolmet Barendregt; Mathilde M. Bekker; Don G. Bouwhuis; Esther Baauw

* Faculty of Industrial Design, † Faculty of Technology Management, Eindhoven University of Technology, Eindhoven, The Netherlands

To cite this Article Barendregt, Wolmet , Bekker, Mathilde M. , Bouwhuis, Don G. and Baauw, Esther(2007) 'Predicting effectiveness of children participants in user testing based on personality characteristics', Behaviour & Information Technology, 26: 2, 133 — 147

To link to this Article DOI: 10.1080/01449290500330372
URL: http://dx.doi.org/10.1080/01449290500330372
Predicting effectiveness of children participants in user testing based on personality characteristics

WOLMET BARENDREGT*†, MATHILDE M. BEKKER†, DON G. BOUWHUIS‡ and ESTHER BAAUW†

†Faculty of Industrial Design
‡Faculty of Technology Management, Eindhoven University of Technology, Eindhoven, The Netherlands

This paper describes an experiment 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 does not lead to finding an unrepresentative sample of the product’s problems.

Keywords: Children; Test participants; User testing; Personality characteristics

1. Introduction

Testing of products, with representative users, is one of the most important aspects of user-centred design. Therefore, it seems quite logical that children should be included as test participants when products intended for this user group are tested. Unfortunately, user testing with children, especially young ones, is not straightforward and requires special attention (Hanna et al. 1997). This is probably one of the reasons why most user testing for children’s products is still done by adult experts (Buckeleiner 1999). However, ‘it is not easy for an adult to step into a child’s world’ (Druin 1999) and therefore involving children in user testing is highly desirable.

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 (Nielsen 2003). However, one problem with including young children in a user test is the fact that children can have difficulty verbalising their thoughts (Boren and Ramey 2000). They often forget to think aloud and need to be prompted 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). A possible solution is to ask children to talk about what they are doing, but to refrain from prompting when they forget to talk. 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.

During various user tests with children on a range of different products it was observed that there are large differences between children 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. Why are some children more revealing during a user test than others? 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, for the evaluation of game concepts, Hanna et al. (2004) selected children who were characterised by their parents as not shy. If it could be possible to predict more precisely which children will make effective participants based on these characteristics, one could do much more cost-effective high-quality testing with these children. Of course, it is essential that these children do not find radically fewer or different problems than the children who make less good participants.

In this article, a study 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. The remainder of this article is divided into six sections. First, 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 problems found by the user group after selection and an example of the effects of selecting the most promising children. Finally, the generalisability of the results is discussed and conclusions are drawn.

2. User test outcome and personality characteristics

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 can 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 (1).

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

(1)

Where \# 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 \# all problems(i) is the total number of problems found by the evaluators in the test session with child i.

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 5 and 13 years old. Blikvanger 5-13 has shown to have ‘good’ to ‘very good’ reliability in terms of internal consistency of the main scales and subscales (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 eight 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 1.

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 $+1$ $z$-scores of the distribution. Examples of the two graphs belonging to one individual profile of a child are given in figures 1 and 2.

2.3 Hypotheses

Based on earlier experiences with children during user tests in the lab, two main hypotheses were formulated. These hypotheses relate to the two measures discussed in section 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:

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 Markopoulos (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 article 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.

The third hypothesis is based on the following definition of Autonomy in the Blikvanger: 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.

2.4 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. 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.

2.5 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 some serious problems to remain undetected (Law and Hvannberg 2004). 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 2003). Subsequently, section 5.3 describes how many problems would have been found if only the most promising children had been used and what the severity of these problems would have been.
3. Method

3.1 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 the minimum wage. All children were between five and seven years old (mean age was 84 months, SD = 5.6 months), nine girls and 17 boys. They were recruited by addressing 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, and the test sessions of three other children were not videotaped correctly. Therefore, the data on these four children were later discarded, leaving 26 children still in the experiment.

3.2 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 2002). 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. Furthermore, a large number of problems were anticipated for children playing this game alone, because even the adult researchers had some problems playing it. This would make the game quite suitable for the experiment.

3.3 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. Generally, the test facilitator did not remind the child to talk aloud during the test. 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 25 minutes the facilitator would tell the child that the test session was over, but that s/he 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.

(Children’s decisions whether to continue the game were gathered for another research goal and will therefore not be discussed here.) Each test session was videotaped, recording a split-screen shot of the face of the child and the onscreen actions.

4. Analysis of the user tests

4.1 Introduction

Recently, several studies have shown that different evaluators often come up with different results (Molich et al. 1998), even when they analyse video tapes of the same user test sessions (Jacobsen et al. 1998, Jacobsen 1999, Vermeeren et al. 2003). This effect is called the ‘Evaluator effect’. Several suggestions have been made to minimise (though probably not eliminate) the evaluator effect. One suggestion is to add more evaluators (Jacobsen 1999), another suggestion is to structure the analysis process (Cockton and Lavery 1999, Vermeeren et al. 2002). A combination of these suggestions might seem an ideal solution. However, when the complexity of the analysis process is increased it also takes much more time and this could make it impossible to find more than one evaluator. For example, in Vermeeren et al. (2002), ratios of session time/ analysis time vary between 1:25 to 1:29, which makes doing the evaluations in this rigorous way a very time-consuming task. For the analysis of the test sessions in the present experiment, it was decided to use a practical combination of these suggestions, which is described in the next subsection.

4.2 The analysis procedure

Many structured data-analysis procedures like, for example 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 this experiment an approach similar to that used by Vermeeren et al. (2003) and Jacobsen et al. (1998) was chosen. In this approach the emphasis is on the second phase of analysing the interaction. Of this first stage the transcription of verbal utterances was applied. An example of this transcription is given in table 2.

For the second stage, Noldus’ The Observer Pro (Noldus 2002) was used, a software package for observational research. With this software observations can be logged with the digital video data. The evaluator just has to click the appropriate behavioural category. 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. As 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 belong to the same problem, ‘Unclear which button is used to shoot the spaceship’.

Following Jacobsen’s advice (1999) to include more than one evaluator, two evaluators were used for the first couple of user tests and for at least one out of five remaining user tests in order to check for ‘coder-drift’ (a change over time in how information is coded). 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 one user test. Both indications belong to the same problem, ‘Unclear which button is used to shoot the spaceship’.

Following Jacobsen’s advice (1999) to include more than one evaluator, two evaluators were used for the first couple of user tests and for at least one out of five remaining user tests in order to check for ‘coder-drift’ (a change over time in how information is coded). 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 one user test. They had to come up first 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 format 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 adapted because it had to be used for an empirical evaluation instead of a heuristic evaluation. The main change was the addition of a section where the evaluators had to provide the explicit link from each problem to the breakdown indications it was based upon. A schematic overview of the analysis process with two evaluators is given in figure 3.

4.3 The breakdown indication types checklist

The DEVAN checklist of breakdown indication types (Vermeeren et al. 2002) is one of the most detailed checklists of breakdown indications available, but it was not created, specifically, to be used for games or with children. Games differ in many ways from other products. For example, they are usually not task-based, and 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). After trying out this original list on collected video material of children playing many different games, some slight adaptations were made. These adaptations were:

1. Due to the exploratory nature of games, the breakdown indication types ‘Correction’ (CORR) and ‘Discontinues action’ (DISC) and ‘Repeated action’ (REP) of the DEVAN checklist were omitted.
2. Because pacing and challenge are very important in games (Pagulayan et al. 2003), the breakdown indication types ‘Impatience’ (IMP) and ‘Bored’ (BOR) were added.
3. The ‘Passive’ (PAS) breakdown indication type was added because some children exhibit this behaviour.
when they don’t know how to proceed without actually showing puzzlement.

4. Due to the used protocol according to which the facilitator would sometimes give help, the breakdown indication type ‘Facilitator provides help’ (HLP) was added.

5. The ‘Searches for function’ (SEARCH) breakdown indication type was omitted because in trials with the coding scheme the difference with ‘Puzzled’ (PUZZ) was unclear to the evaluators.

6. The ‘Execution difficulty’ (DIFF) breakdown indication type was collapsed with the ‘Doubt surprise frustration’ (DSF) breakdown indication type because they all occur directly after the child has executed, or has tried to execute, an action.

The final list of breakdown indication types still looks very much like the one used in DEVAN and is depicted in table 3.

Two evaluators unfamiliar with this study tested the reliability of the final list of breakdown indication types on a predefined set of 29 breakdown indications for another game called ‘Rainbow, the most beautiful fish in the ocean’. Of the 29 given breakdown indications, 26 were coded as the same breakdown indication type, resulting in a kappa of 0.87.

4.4 Problems in games

Naturally, some aspects of the game that seem to be problematic are actually part of the fun. For example, it can be quite hard to shoot spaceships when they fly very quickly, resulting in breakdown indications of the type ‘Execution problem’ or ‘Doubt surprise frustration’. The decision whether a breakdown indication should be considered an actual problem was left to the judgement of the evaluators. When a breakdown indication was not 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.

4.5 Reliability analysis

Due to concern for the evaluator effect, eight out of the 26 user tests were analysed 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 as proposed by Hertzum and Jacobsen (2001):

$$\frac{|P_1 \cap P_2|}{|P_1 \cup P_2|}$$  \hspace{1cm} (2)

In this equation, \(P_1\) and \(P_2\) are the sets of problem indications detected by evaluator 1 and 2, respectively. As it is practically impossible for two evaluators to score behaviour at exactly the same time, a time frame of two seconds before and after the actual time stamp was allowed to determine (in)equality of two problem indications.

The average any-two agreement measure for these eight analysed user tests 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).

4.6 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.

4.7 Determining representativeness

As described in section 2.5, 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 two severity measures: Frequency Severity and Impact Severity. Subsequently, the problems that are missed by a group of most-promising children are discussed in terms of both types of 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
Table 3. Definition of 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>Subgame stopped</td>
<td>The user stops the subgame 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: 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 frustrated the user.</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>PUZ</td>
<td>Puzzled</td>
<td>The user indicates: 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>PER</td>
<td>Perception problem</td>
<td>Recognition of error or misunderstanding: the user indicates to recognise a preceding error or misunderstanding.</td>
</tr>
<tr>
<td>BOR</td>
<td>Bored</td>
<td>The user indicates not being able to hear or see something clearly.</td>
</tr>
<tr>
<td>RAN</td>
<td>Random actions</td>
<td>The user verbally indicates being bored.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The user non-verbally indicates being bored by sighing or yawning.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The user indicates verbally or nonverbally to perform random actions.</td>
</tr>
</tbody>
</table>

determine criticality of problems (Rubin 1994, Nielsen 2003). Nielsen (2003) defines only two levels for each type of severity, which gives a rather crude classification, while Rubin (1994) defines four levels of severity. For the classification in this experiment it was decided to define three levels of severity for both types of severity. There are two reasons for this decision. The first reason is that the children are allowed to explore freely, making it unlikely that they all visit exactly the same screens within the game. Therefore, Rubin’s highest category of Frequency Severity would not contain many problems. The second reason is the fact that the computer game used in the user tests was already commercially available. Therefore, Rubin’s highest category of Impact Severity would probably not contain many problems.

In practice, both measures will often be used in combination to determine the overall severity of each problem. However, they will be discussed separately in this article to provide a clear overview of the situation.
4.7.1 Frequency Severity. The Frequency Severity classification of each problem was determined as follows:

- High frequency severity: Problem was experienced by \( \geq 38\% \) of the children.
- Average frequency severity: Problem was experienced by \( 20–37\% \) of the children.
- Low frequency severity: Problem was experienced by \( \leq 19\% \) of the children.

4.7.2 Impact Severity. To take into account the fact that the impact of a problem can differ from person to person, the Impact Severity for each problem was determined as follows:

- For each child that experienced the problem a score of 0 was given if the child could continue without further help,
- A score of 1 was given if the child could continue with a little help,
- A score of 2 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 2.0, the lowest possible score 0.0. For example, suppose two children experienced a certain problem. 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 0, for the second child the impact severity would be 1. The overall Impact Severity of this problem would then be 0.5.

The Impact Severity classification of each problem was determined as follows:

- High impact severity: \( 1.4 \leq \text{Overall Impact Severity} \leq 2.0 \)
- Average impact severity: \( 0.7 \leq \text{Overall Impact Severity} < 1.4 \)
- Low impact severity: \( \text{Overall Impact Severity} \leq 0.7 \)

4.7.3 Selection of test subjects. To judge whether selection of the ‘most-promising’ children would have resulted in missing an unexpectedly high percentage of problems or many high-severity problems it should be decided how many, and which, children would have participated if only the ‘most-promising’ ones were selected. By comparing the problems uncovered by these children with the problems uncovered by the whole group, it is possible to get an indication of the representativeness of the problems of this subgroup.

Using the formula \( 1 - (1 - p)^n \), where \( n \) is the number of test participants and \( p \) is the detection rate of a given problem, several researchers have shown that the first 3–5 test participants are enough to find 80 per cent of the usability problems (Virzi 1992, Nielsen 1994). This means that the average detection rate \( p \) of a problem should be as high as 0.42 with only three test participants, and that it should be 0.28 with five participants in a test. However, detection rates are often much lower (Lewis 1994, Bekker et al. 2004) and in that case many more test participants are needed to uncover 80 per cent of all problems. The discussion of representativeness of the problems found by a group of most-promising children will use a group size that, according to this formula, should find 80 per cent of all problems when using the actual \( p \)-value of this user test.

5. Results

5.1 Hypotheses

Linear regression analyses were performed to test the hypotheses. The first hypothesis asserts that:

\[ \text{H1: There is a significant positive correlation between the score on Curiosity and the number of problems.} \]

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:

\[ \text{H2: There is a significant correlation between the scores on Extraversion and Friendliness and the proportion of problems indicated through self-initiated spoken output.} \]

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 for any of the 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:

\[ \text{H3: There is a significant negative correlation between the score on Autonomy and the proportion of the problems indicated through self-initiated spoken output.} \]
5.2 Exploring other factors

5.2.1 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 4. The results for the proportion of problems indicated through self-initiated spoken output are given in table 5.

The personality characteristic Intelligence alone 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 possible 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 could give 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.

Apart from the hypothesised main scale predictors Extraversion and Friendliness, no alternative predictor or set of predictors for the proportion of problems indicated through self-initiated spoken output was found.

5.2.2 Subscales. For the 16 subscales 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

<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>x</td>
<td>x</td>
<td>x</td>
<td>x**</td>
<td>7.048</td>
<td>0.227</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>3.290</td>
<td>0.121</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>0.653</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>0.528</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>0.059</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x**</td>
<td>6.453</td>
<td>0.359</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>4.415</td>
<td>0.277</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>4.372</td>
<td>0.275</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>3.789</td>
<td>0.248</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>3.716</td>
<td>0.244</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.584</td>
<td>0.121</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>0.872</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>1.878</td>
<td>0.066</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>0.467</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>0.403</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x**</td>
<td>4.870</td>
<td>0.399</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>4.643</td>
<td>0.388</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>4.558</td>
<td>0.383</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>4.430</td>
<td>0.377</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>3.202</td>
<td>0.304</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>2.976</td>
<td>0.289</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>2.399</td>
<td>0.247</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>2.379</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.343</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.744</td>
<td>0.092</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>4.431</td>
<td>0.458</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>3.695</td>
<td>0.413</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>3.545</td>
<td>0.403</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>3.434</td>
<td>0.395</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>1.795</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>3.436</td>
<td>0.462</td>
</tr>
</tbody>
</table>

*p < 0.05, **p < 0.01.
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 indicate 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, but the subscale predictor Dominance could be investigated further.

Backward regression indicates a much larger set of predictors, containing Self-Confidence, Approach seeking, Curiosity, Emotional stability, Manageability ($df = 25$, $F = 5.011$, $R^2 = 0.556$, $p = 0.004$).

For the proportion of problems indicated through self-initiated spoken output, both the stepwise and the forward regression analysis indicates a combination of Approach seeking and Altruism as a set of predictors ($df = 25$, $F = 4.874$, $R^2 = 0.298$, $p = 0.017$). Since Approach seeking 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, Approach seeking, Affect, Emotional stability, Manageability, Dominance and Positive emotionality ($df = 25$, $F = 3.083$, $R^2 = 0.545$, $p = 0.026$). Altruism and Approach seeking are the only predictors present in all analysis results.

The predictors indicated by the backward regressions could all be investigated further.

### 5.3 Representativeness

To determine the representativeness of problems found by a selection of children it should first be decided how many children should have participated in order to find 80 per cent of the problems. The detection rate in this experiment was
calculated to be as small as 0.14. Using the formula $1 - (1 - p)^n$ it can be determined that 11 children should have been selected to find 80 per cent 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 the standard score on Curiosity. Subsequently the children were ranked according to the Extraversion/Friendliness combination standard score. Based on the regression equation, the combination score was calculated by subtracting the standard score for Friendliness from the standard 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.

5.3.1 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 per cent. All 26 children together found 109 problems. The group of 11 most-promising children would have found 82 problems, which is 75 per cent of all problems. This is not far from the expected percentage.

5.3.2 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 per cent of the problems. Of the 109 problems found by all children, 11 were classified as high frequency severity problems, 20 as average 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 per cent of the high frequency severity problems, 100 per cent of the average frequency severity problems, and 65 per cent of the low frequency severity problems.

Of all 109 problems, six were classified as high impact severity problems, 18 as average 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 per cent of the high impact severity problems, 88 per cent of the average impact severity problems, and 72 per cent 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 average 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, does not damage the representativeness of the results.

5.4 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 5.2.

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 been indicated 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 6.

Regarding Frequency Severity, the group of least-promising children would have found 100 per cent of both the average and high frequency severity problems. Of the 78 low frequency severity problems, they would have found only 58 per cent. 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 7.

Regarding Impact Severity, the group of least-promising children would have found 83 per cent of both the high impact severity problems. Of the 18 average impact severity problems, they would have only found 61 per cent. Of the low impact severity problems they would have found 71 per cent. 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 8.

Table 6. 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>
<tr>
<td>Number of verbalised problems</td>
<td>43</td>
<td>28</td>
</tr>
<tr>
<td>Average number of children finding a problem*</td>
<td>3.5</td>
<td>2.8*</td>
</tr>
</tbody>
</table>

*Over 44 problems found by both groups.

*p < 0.001, one-tailed.
Table 7. Numbers of high, average 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 group of least-promising children</th>
<th># Problems found by group of most-promising children</th>
<th># Problems found by all 26 children</th>
</tr>
</thead>
<tbody>
<tr>
<td>High severity</td>
<td>11 (100%)</td>
<td>11 (100%)</td>
<td>63% (100%)</td>
</tr>
<tr>
<td>Average severity</td>
<td>20 (100%)</td>
<td>20 (100%)</td>
<td>61% (100%)</td>
</tr>
<tr>
<td>Low severity</td>
<td>78 (63%)</td>
<td>51 (63%)</td>
<td>58% (100%)</td>
</tr>
</tbody>
</table>

Table 8. Numbers of high, average 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 severity category</th>
<th># Problems found by group of least-promising children</th>
<th># Problems found by group of most-promising children</th>
<th># Problems found by all 26 children</th>
</tr>
</thead>
<tbody>
<tr>
<td>High severity</td>
<td>6 (83%)</td>
<td>5 (83%)</td>
<td>72% (83%)</td>
</tr>
<tr>
<td>Average severity</td>
<td>18 (88%)</td>
<td>16 (88%)</td>
<td>61% (88%)</td>
</tr>
<tr>
<td>Low severity</td>
<td>85 (72%)</td>
<td>61 (72%)</td>
<td>71% (85%)</td>
</tr>
</tbody>
</table>

This 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 average 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.

6. Discussion

6.1 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 paper, 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 study in this paper indicates that extraversion should not be considered as an 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 an experimental foundation for the common practice of using children who are characterised as not shy for evaluation purposes, like the 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.

6.2 Generalisability

The results of this experiment are based on only 26 children playing one game. Further research has to be done to determine whether the results also hold for other situations with other groups of children. There is, however, an important indication that the results will also hold for different games. In another pilot study employing seven children, each played the same two adventure type games (two other games than ‘Milo and the Magical stones’). The parents had filled in the Blikvanger questionnaire prior to the test. After analysing the user tests in the same way as described in section 4.2 and ordering the children 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.

Although strictly speaking the results described here apply only to the selection of children based on the Dutch personality test ‘Blikvanger’, the general acceptance of the underlying concepts – the Big Five personality traits – makes it possible to obtain the same results using other tests. For the selection of adult test users the results could still hold if the same protocol of voluntary talking aloud could be used. However, for adults, usually the strict think-aloud protocol is followed according to which they are reminded to keep on talking. As 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.

6.3 Practical advice

As was shown in section 5.3.2, selecting children based on the personality characteristics, Extraversion, Friendliness and Curiosity, does result in finding almost all high and average severity problems. The advice for evaluation practitioners would be to make a selection of children based on the personality characteristics, Extraversion, Friendliness and Curiosity. Especially in commercial environments, where often, smaller numbers of test participants are used, the benefits of choosing ‘most-promising’ children may be much bigger than in this experiment.

Preferably these ‘most-promising’ 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.

7. Conclusion and future work

The experiment described in this article 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. Using these children in a user test can optimise testing efforts. 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 subscale 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.

The interests of our project mainly concern the development and comparison of evaluation methods for young children. We are planning to use the results of this experiment, primarily to look more closely into the effects of changing the way the facilitator and the children interact. For example, it could be possible that children who are less good at self-initiated spoken output can be persuaded to do this more often when the facilitator behaves differently towards them. This further research hopefully leads to even more insights in how to improve the results of user testing with children.

Acknowledgements

We would like to thank Silvia Crombeen and Mariëlle Biesheuvel for conducting the user tests. We would also like to thank the children and teachers of primary school de Bremsbucht for participating in our research. Furthermore, we would like to thank Prof. Dr. G. Keren, Prof. Dr. C. Snijders and three anonymous reviewers of this paper for their constructive advice. The Innovation-Oriented Research Program Human-Machine Interaction (IOP-MMI) of the Dutch government supported this research.

References


RUBIN, J., 1994, Handbook of usability testing: how to plan, design, and conduct effective tests (Chichester: Wiley & Sons).