A Method to Evaluate Disabled User Interaction: A Case Study with Down Syndrome Children

Isys Macedo and Daniela G. Trevisan

Universidade Federal Fluminense, IC – Instituto de Computação
Rua Passo da Pátria 156, Bloco E 24210-240 Rio de Janeiro, Brasil
{isysmacedo,daniela}@ic.uff.br

Abstract. Testing products with representative users is a key factor for user-centered design. When such representative users are disabled children the user testing process becomes a challenge and in this case evaluation methods based on heuristics and inspection could not attend the final user needs. The major purpose of our research is to provide an evaluation method that could measure disabled children interaction. This work first discusses the development of the coding scheme based on the detailed video analysis method which was adapted to observe interaction of children with Down syndrome. After that we demonstrate the method reliability by applying the cohen's kappa coefficient and the any-two agreement measure. Finally we discuss how this method could be used to evaluate usability and fun problems.

Keywords: Usability game evaluation, children interaction, down-syndrome.

1 Introduction

Recently the concern with the development of people with special needs and their inclusion in society has grown. The first step to make this possible is to provide children with special needs ways of stimulus that are pleasant and efficient at the same time. One way to combine these two features is through digital games that stimulate most of the speech, comprehension, attention, perception and other factors needed for a good social life.

The learning process of children with Down syndrome occurs at a slower rate compared to other children of the same age, for this reason they take longer to learn how to read, write, do math, among other tasks. [5] Moreover, these children are more likely to have more interaction problems that must be taken into account when someone proposes to develop some kind of software for this specific audience. In this case, evaluation methods of interaction based on inspection and heuristics cannot meet this need since they do not involve the end user.

Due to the specific needs of children with Down syndrome, evaluators may not predict usability and fun problems and when computer games are evaluated it is important to fix both. In spite of that, there is no coding scheme of behavior that indicates these problems in computer games for children with any kind of special need. In this way, this article proposes a new coding scheme to detect usability and fun

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2013, Part I, LNCS 8009, pp. 50–58, 2013.

problems in games developed for children with Down syndrome. The method is based on a list of breakdown indication types to evaluate children's computer games [2]. Nevertheless, the definitions of existing breakdown indications probably need to be changed, new breakdown indications need to be added, and some indications have to be removed. The article starts with a brief introduction to the method created by Barendregt and Bekker [2] and then immediately describes the changes made to adapt it to children with Down syndrome. After that, it will be described how user interaction was captured, application of the method and also the results that prove the reliability of the method. Finally, the article discusses the results for reliability of the method and also how this method could be combined with other in order to better study the fun criterion.

2 DEVAN

DEVAN [5] is a tool for detailed videos analysis of user test data. It makes use of a table format for representing an interaction at multiple levels of abstraction.

DEVAN method was originally designed to detect usability problems in products targeted at adults and it was adapted by Barendregt and Bekker [2] to assess usability and fun problems in games targeted at children.

2.1 DEVAN Adapted for Children

Based on the list of breakdown indications of the DEVAN method [5], Barendregt and Bekker [2] presented a new list of breakdown indications that reflects the behavior observed in children when they indicate problems of fun and usability.

Despite the fact that the DEVAN method was developed to detect usability problems in task-based products for adults, for the new method [2] it was chosen not to use explicit tasks. The goals set by the tasks may interfere with the goals provided by the game, because children feel obligated to fulfill the tasks and also to achieve the goals of the game [13].

To enable detection of fun problems, the taxonomy defined by Malone and Lepper in 1987 [12] was explored. This taxonomy consists of four main heuristics: challenge, fantasy, curiosity and control. From each heuristic the following indications were added to the list: *help*, *bored*, *impatience* and *dislike*.

Moreover, in [2] Barendregt and Bekker observed the need to include other breakdown indications. Usually, games with texts that are difficult to read or with complex verbal explanations generate attention problems, so the breakdown indication *perception problem* was added. Sometimes children remain passive when an action is expected because they don't know how to proceed. To represent these situations *passivity* was included.

Finally, the indication *wrong action* has been added to include situations in which the child does not understand how the game works properly and when asked about a particular action, the answer is not correct.

2.2 Adaptations in the Coding Scheme

In order to meet the need of evaluating digital games for children with Down syndrome, it was necessary to remove, add or redefine meanings of some information contained in the list of previous work [2].

Removed Breakdown Indications. Basically children who have Down syndrome have more difficulty expressing themselves verbally. In [2] there were breakdown indications which were only based on verbal explanations. In this work it could be difficult to observe these and probably its frequency would be very low. Because of that the following breakdown indications were removed: wrong goal, wrong explanation, recognition, 'doubt, surprise and frustration'.

Maintained Breakdown Indications. Six breakdown indications from [2] were maintained with no changes in their definition. The indications and reasons for keeping them are presented below. *Wrong action* was kept on the list because it can be observed when a child clicks on a non-clickable area or performs an action that was not expected at that moment. The difficulty of interacting with a physical game device is characterized by the indication *execution problems*.

If the game responds slowly, or if the user fails to perform some type of command in the game, the indication *impatience* appears when the user repeatedly click on a button or make more abrupt movements in an attempt to do something work or get a faster response. Games for children with special needs should also entertain their users and propose challenges. But when the challenge proposed is very easy and does not interest the child to continue playing, certainly the indication *subgame stopped* will be noticed. When the game fails to stimulate the user's curiosity, the child begins to yawns and sighs which are a clear demonstration of the indication *bored*. The game interface should be able to drive the user in a simple and efficient way, if this does not happen, interaction problems can occur and the user can get confused, pointing the indication *puzzled*.

Modified Breakdown Indications. Five breakdown indications from [2] were maintained but had their definitions changed. As previously mentioned children with Down syndrome have difficulty to verbalize their feelings and thoughts, for this reason the need of verbalization was removed from the definition of indications: *perception problem* and *random actions*.

Difficult to achieve goals, hinder the child to proceed without the intervention of a mediator. Mediators often help children with Down syndrome when they realize that the child cannot proceed, i.e. without a verbal request, thereby avoiding serious problems. *Help* also had its definition changed, again precluding the need for verbalization. If the fantasy provided by the game is too childish or too scary the child may express *dislike*, which can be observed only through facial expressions.

If the child does not know how to perform some action he or she tends to be *passive* just staring at the computer screen. But there are other reasons that could cause the child to have that reaction, for example, he or she did not want to perform an activity because it is difficult or boring, or even by lack of stimulus.

New Breakdown Indication. During an interaction session it was observed that one child performed a wrong action just to enjoy the reaction of a character. Such behavior did not fit perfectly into *wrong action* indication, appearing then the need to insert an indication of *intentional wrong action*. In this action the child knows that it is not the correct action, but he or she is still doing it only for the purpose of fun.

3 Capturing the User Interaction

Our study case is based on the user interaction with JECRIPE – a game that has a purpose of stimulating preschool children with Down syndrome [1]. Four children with Down syndrome, aged between 6 and 12 years (mean = 9 years), joined the group of users (Figure 1). None of the children had experienced JECRIPE before and was chosen not to use explicit tasks. From the interaction session we obtained a video of approximately 20 minutes containing interactions of children with the three game scenarios (Figures 2 (b), (c) and (d)). Ethical procedures regarding the user participation in the video was also performed.

Fig. 1. Capturing the user interaction in a video recording

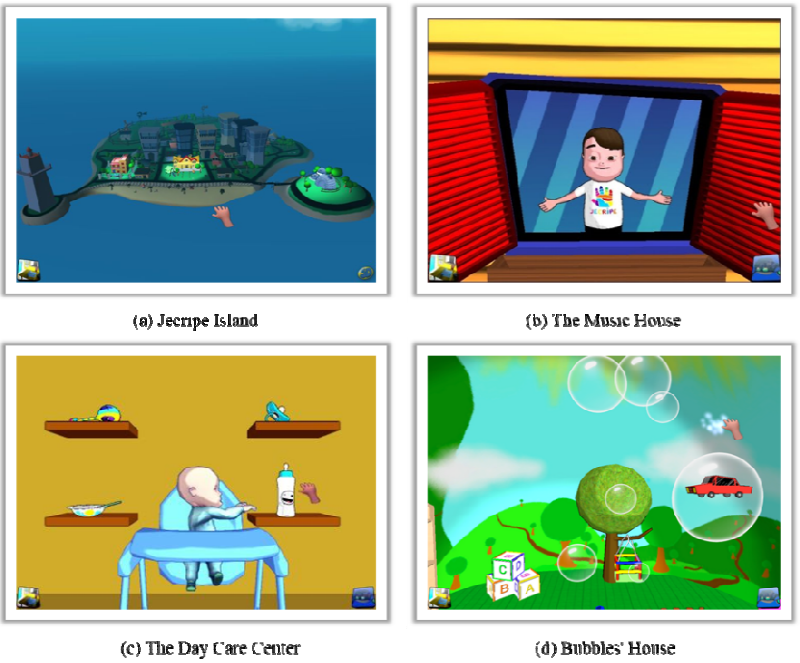

Fig. 2. JECRIPE interactive scenarios

4 Coding the User Interaction

To encode the video, we invited eleven students from the Human Computer Interface class that had little or no experience in such evaluations. They were divided into 4 groups. The training session consisted of a 30 minute presentation in which videos exemplifying the codes to be used were shown (see Appendix).

Evaluators were instructed to observe the occurrence of problems and to examine whether these problems characterized some of the indications present in the table. Each interaction problem detected and the corresponding code should be noted the instant of its occurrence. The average time for analysis and coding was approximately 43 minutes.

5 Results of the Method Reliability

It is most unlikely that different evaluators will agree exactly, by giving the identical result for all evaluation sessions. The any-two agreement method measures the extent of agreement on what problems the system contains for pairs of evaluators [3]. For each comparison the number of agreements, disagreements and single points were recorded, always considering the margin of 4 seconds to be counted as the same observation point (see Figure 3).

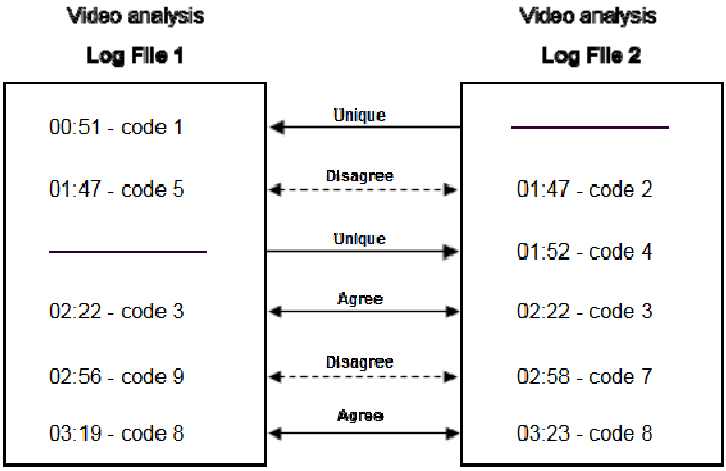

Fig. 3. Video coding analysis

The results of the comparisons for each pair of evaluations are shown in Table 1. The average score was 44%, exceeding the average of 38.5% obtained in [2]. The number of single points was the factor that affected most negatively the percentage obtained. It happened mainly due to the fact that some groups of evaluators noted long intervals while others noted small intervals for the same interaction problem observed. The points of disagreement were investigated in order to determine the indications that generated more conflicts.

The codes *wrong action* and *execution problems* were the most confusing. In the video analyzed, children often could not perform an action especially because of the difficulty in handling the mouse and on situations like that a group classified it as *wrong action* and another as *execution problems*. This divergence in classification may have occurred because the evaluators may have interpreted differently the definitions of the codes or they had different focuses on that moment. For example, while one group was more attentive to perform the sequence of actions in the game, the other watched not only the realization of the actions but also the interaction with the physical devices, in this case the mouse, the game.

Table 1. Results for each any-two comparison

Evaluation **A** x Evaluation **B**	**Any-Two (%)**	**Agreements**	**Disagreements**	**Unique A**	**Unique B**
Evaluation **1** x Evaluation **2**	49	39	4	29	7
Evaluation **1** x Evaluation **3**	47	49	7	18	31
Evaluation **1** x Evaluation **4**	44	38	17	19	12
Evaluation **2** x Evaluation **3**	39	36	7	6	44
Evaluation **2** x Evaluation **4**	45	33	7	9	25
Evaluation **3** x Evaluation **4**	41	40	16	30	12

Another index used to measure the method consistency was the Cohen's Kappa [4], for interpreting the result the following guideline was used [11]:

- Less than 40% = low agreement.
- Between 40% and 60% = average agreement.
- Between 60% and 75% = good agreement.
- More than 75% = excellent or perfect agreement.

For that purpose two new evaluators received the same training session described before and a list of observation points for which they individually had to pick a code.

This list was created by taking the list of all four groups of evaluators in the first evaluation. When at least three of the four groups of evaluators in the first experiment agreed on an observation point (but not necessarily on the code), it was included in the list of observation points, resulting in a list of 47 fixed observation points. Of all the 47 points contained in the list, 30 were also encoded in both assessments, producing a concordance of approximately 64%. Sometimes it was noted that the discrepancy occurred due to the priority order, e.g. an evaluator puts a code as the most important while the other placed it as the second.

Table 2 shows a comparison between the percentages achieved in any-two agreement and also Cohen's kappa for this work and [2]. It is important to emphasize that the evaluators considered in this study for the validation of Cohen's Kappa have different levels of experience in usability evaluation (one without experience and the other an experienced evaluator) and they encoded a list of 47 points. Whereas the evaluators in [2] codified a list of 26 points.

Table 2. Comparison between the results of the two methods

	Any-two agreement		*Cohen's Kappa*	
	Percentage	**Number of evaluators**	**Percentage**	**Fixed Points**
Barendregt & Bekker (2006)	38.5%	4	92%	26
This work	44%	11 (divided into 4 groups)	64%	47

6 Final Considerations

Finally we conclude that such adaptations performed in the DEVAN method were suitable to evaluate interaction of children with Down syndrome.

Two measures have been implemented to verify the reliability of the method. Four groups of evaluators participated in the first measure used as part of the validation of the method, and then two new evaluators collaborated to compute the second measure. The result for each measure was satisfactory and the combination of these proves the reliability of the method.

Note that due to lack of experience of evaluators who participated in the first step of method validation (any-two agreement), it can be stated that the method has an easy application that its indications and definitions are clear. Based on the result of Cohen's Kappa (64% agreement) that included the participation of evaluators with

distinct experience levels, we can assume that even inexperienced evaluators will be able to apply the method after receiving an appropriate training. However both assumptions need to be further investigated.

As future work, we could combine the method with other previous investigations in which the user informs more directly how much fun he had playing the game. For example, by using the Smileyometer [9] it would be possible to verify the impact caused by the fault of the game as the criterion of enjoyment reported by the user as well as where the indications of the DEVAN method, are appropriate and effective to evaluate the fun criterion.

Based on the statement made by [10] that children between 7 and 8 years of age are able to understand and distinguish the concepts of usability, fun and potential for learning, we believe that the methods reported by [9] may also be suitable for children of the same age range of the participants in this study, because the methods are not composed of complex questionnaires and they are also easy to apply.

Besides that we intend to extend this method to estimate users interactions with others kinds of disabilities and also link the detected usability problems with appropriate design guidelines as those pointed by [8].

References

1. Brandão, A., et al.: JECRIPE: stimulating cognitive abilities of children with Down Syndrome in pre-scholar age using a game approach. In: Proceedings of the 7th International Conference on Advances in Computer Entertainment Technology (ACE 2010), pp. 15–18. ACM, New York (2010)
2. Barendregt, W., Bekker, M.M.: Developing a coding scheme for detecting usability and fun problems in computer games for young children. Behavior Research Methods 38(3), 382–389 (2006)
3. Hertzum, M., Ebbe, J.N.: The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods. International Journal of Human-Computer Interaction 15(1), 183–204 (2003)
4. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics 22(2), 249–254 (1996)
5. Vermeeren, A.P.O.S., den Bouwmeester, K., Aasman, J., de Ridder, H.: DEVAN: A detailed video analysis of user test data. Behaviour & Information Technology 21, 403–423 (2002)
6. Brandão, A., Trevisan, D.G., Brandão, L., Moreira, B., Nascimento, G., Vasconcelos, C.N., Clua, E., Mourão, P.: Semiotic Inspection of a Game for Children with Down Syndrome. In: Proceedings of the Brazilian Symposium on Games and Digital Entertainment (SBGAMES 2010), pp. 199–210. IEEE Computer Society, Washington, DC (2010)
7. Dix, A., Finlay, J., Abowd, G.D., Beale, R.: Human-Computer Interaction, 3rd edn. Pearson (2004)
8. Jokisuu, E., Langdon, P., Clarkson, P.J.: Modelling Cognitive Impairment to Improve Universal Access. In: Stephanidis, C. (ed.) Universal Access in HCI, Part II, HCII 2011. LNCS, vol. 6766, pp. 42–50. Springer, Heidelberg (2011)
9. Read, J.C., Macfarlane, S.J., Casey, C.: Endurability, Engagement and Expectations: Measuring Children's Fun. In: Proceedings of the Interaction Design and Children, pp. 189–198. Shaker Publishing, Germany (2002)

10. Macfarlane, S., Sim, G., Horton, M.: Assessing usability and fun in educational software. In: Proceedings of the 4th Conference on Interaction Design and Children (IDC 2005), pp. 103–109. ACM, New York (2005)
11. Robson, C.: Real World Research: A resource for social scientists and practitioner researchers. Blackwell Publishers, Malden (1993)
12. Malone, T.W., Lepper, M.R.: Making learning fun: A taxonomy of intrinsic motivations for learning. In: Snow, R.E., Farr, M.J. (eds.) Aptitude, Learning, and Interaction III: Cognitive and Affective Process Analysis, pp. 223–253. Erlbaum, Hillsdale (1987)
13. Barendregt, W., Bekker, M.M., Speerstra, M.: Empirical evaluation of usability and fun in computer games for children. In: Rauterberg, M., Menozzi, M., Wesson, J. (eds.) Proceedings of the IFIP 8th International Conference on Human–Computer Interaction, pp. 705–708. IOS Press, Amsterdam (2003)

Appendix

DEVAN Method for children with Down syndrome.

Code	Description	Definition
Breakdown Indication Types Based on Observed Actions with the game		
ACE	Wrong Action	An action does not belong in the correct sequence of actions. An action is omitted from the sequence. An action within a sequence is replaced by another action. Actions within the sequence are performed in reversed order. The user performs a wrong action unintentionally.
ACP	Intentional wrong action	The user knows that the action is wrong, but still performs this action only to have fun.
AJU	Help	The user cannot proceed without help or the researcher has to intervene in order to prevent serious problems. The user is helped to do some action.
ANT	Dislike	The user indicates disliking something.
CON	Puzzled	The user indicates not knowing how to proceed.
IMP	Impatience	The user shows impatience by clicking repeatedly on objects that respond slowly, or when it takes too much time to reach the desired goal.
PAS	Passive	The user stops playing and does not perform the expected action.
PEX	Execution Problems	The user has physical problems during interaction with the game. The user has motor skill problem.
PPR	Perception Problem	The user indicates not being able to hear or see something clearly, not understanding how to proceed.
RAN	Random actions	The user performs random actions.
STP	Subgame stopped	The user stops the subgame before reaching the goal.
TED	Bored	The user indicates being bored by sighing or yawning.