Design and empirical testing of a checklist for the evaluation of multimedia software for children

Duda, S.
Humboldt-Universität zu Berlin, Institut für Psychologie
Ingenieurpsychologie
Oranienburgerstr. 18, 10178 Berlin, Germany
Phone +49 (0)30 285165-1
sabrina.duda@rz.hu-berlin.de, http://www.kids-soft.ipfb.de

Abstract

An expert checklist for evaluating children's software has been designed and validated on ten children in the second grade. In the first part of the investigation ten psychology students evaluated three edutainment games, all of different quality, by means of the checklist. Based on these results the items on the checklist were then analyzed and filtered (analysis of variance, item analysis). In this way the length of the checklist was reduced considerably. The remaining checklist items became the basis for calculating a new software evaluation score.

In the second part of the investigation seven and eight-year-old children played with the same edutainment games as the students. While doing so they were observed and subsequently interviewed by the author. A regression analysis was used in predicting the outcome of the observation of and interviews with the children using the newly calculated checklist results. The results show that it is possible to predict children's reactions to certain edutainment games by using the checklist.

Due to the low number of subjects and the use of only three different edutainment games, however, the results should be regarded as a kind of preliminary test indicating a general tendency.

Keywords

Software, multimedia, edutainment, children, evaluation, checklist, usability.

1 INTRODUCTION

More and more, children are recognized as a new group of software users. While softwareergonomics dealt in the past mainly with software for adults, a new tendency is now arising whereby the special characteristics of the designing and usability testing of software for children are examined. Hanna, L., Risden, K., and Alexander, K. J. (1997) of Microsoft have developed guidelines for usability testing with children. Robertson, J. W. (1994) criticizes the lack of attention given to the usability of educational software. She proposes various methods of testing usability with children.

At the CHI-Conference in 1997 there was a special program for children called CHI-Kids (Druin, A., 1997). A Listserv list CHI-Kids exists as well.

For the past two to three years software architects have been attempting to produce educational software designed especially for children and boasting a high quality. The target group of the software industry is becoming younger and younger; there are even CD-ROMs available which are meant for three-year-olds to learn and play with. The new genre is edutainment software: it combines education with entertainment.

Earlier in the history of the industry many software firms blindly launched products onto the market with the hope that the catchword 'multimedia' would suffice to convince the consumer – a hope that was soon shattered; many firms disappeared from the market. Today, quality is becoming an increasingly important factor. 'Kindersoftware-Ratgeber 1998' (Feibel, T., 1997), a guide to children´s software, surveys the vast market. German youth welfare departments publish brochures in which certain computer games for children are recommended. The 'Unterhaltungssoftwareselbstkontrolle' (an organization which oversees entertainment software, hereon referred to as USK) examines computer games and confers to them a rating comparable to the age restrictions imposed by the 'Freiwillige Selbstkontrolle der Filmwirtschaft' (or FSK, which oversees cinematic entertainment) in the motion picture industry. The USK focuses mainly on the problem of violence and pornography in software. The USK rating is commonly accepted. However, it is not to be understood as a recommendation in relation to quality. This fact is often overlooked by parents.

There are institutions and persons responsible for establishing the criteria by which computer games are to be recommended, such as the 'Arbeitsgemeinschaft Kinder- und Jugendschutz', an organization involved in the protection of children and young people (Lerchenmüller-Hilse, H., 1995), or Geisler, T. (1995) or Zey, R. (1994). The most extensive criteria were developed by Fritz, J. and Fehr, W. (1996) of the 'Computerprojekt Köln'. They are used for 'Computerspiele auf dem Prüfstand' (an examination of computer games), a series of booklets published by the 'Bundeszentrale für politische Bildung' (a German federal bureau for political education). These criteria place emphasis on the contents of the games and on the pedagogical aspect pertaining to them.

The 'Institut für Medien und Bildung' (Institute for Media and Education) confers a seal of approval to multimedia software of high didactical value. The criteria by

which this value is measured focus on the educational aims and user-friendliness (source: L.A. Multimedia - Magazin für Medien und Bildung, 1997; no author mentioned).

At present there is no criteria catalogue available to provide a detailed rating of the quality of various edutainment software. But there is a great need for such catalogues and official software ratings. Some software firms are interested as well in having the quality of their products tested.

2 ISSUE

In order to address the need for an evaluation instrument, the purpose of this work has been to develop a checklist for evaluating children´s software. An effort was made to develop a general checklist which can be applied to various types of software. This study deals with edutainment software; that is, software with which children aged four to ten can play and learn. The checklist can be adapted so as to be suitable for purely educational software.

The checklist is designed to be able to differentiate between three edutainment games of different quality. In addition, it is to be validated by means of empirical testing with children. The checklist scores are to correlate with the children´s behavior and their interview scores. Games with a high score on the checklist should therefore be preferred by the children.

3 CHECKLIST DESIGN

A test version of the checklist, with 236 items, was developed.

The aspects peculiar to evaluating software which is neither designed for office work nor meant to serve as a tool for carrying out assignments had to be considered. Edutainment software is a 'tool' for having fun and learning. To achieve both the user has to in fact carry out assignments given in the software, such as finding hidden things, solving problems, or training his sensorimotor abilities; but not in the sense of being an employee. The dimensions of conventional software evaluation as described for instance in EVADIS II (Oppermann, R.; Murchner, B.; Reiterer, H.; Koch, M.; 1992) or the ISONORM 9241/10 checklist (Prümper, J. & Anft, M., 1993) were not suitable for evaluating children´s software and so a new instrument had to be constructed.

A second source of complication was the fact that the users in this case are children. It is not possible to ask a child a large number of detailed questions about such aspects of software as usability. Therefore, adult experts have to evaluate this kind of software and the problem of perspective arises. It had been decided that the adults would judge the software from their own personal point of view. Some (very few) questions require that the adults answer from the perspective of a child. When these questions occur this is explicitly mentioned.

Dimensions of the checklist

- *Cover and booklet*
  Does the cover contain all important information? Is the booklet (a very thin user's manual with information for the parents) well written?
- *Entertainment value*
  In software for children, especially edutainment software, the entertainment value plays a great role. If the software does not entertain, children will not engage themselves in it, and no educational effect will be achieved.
- *Suitability for children*
  The design of the software should take into account the special needs of children. Tasks, for example, should be suitable for children, and the child should be able to identify with the characters.
- *Ease of use*
  Usability is a central factor in children´s software. Without a clear, consistent design, appropriate feedback, and a help function, the joy of playing is spoiled.
- *Load*
  Children´s software should not of course cause stress by pressuring the users to accomplish a task within a very short period of time, or by exhausting their sensorimotor skills.
- *Educational value*
  The educational value is more or less explicit in children´s software. There is software which conveys facts or information usually acquired at school and there is software which implicitly encourages general problem-solving. (And of course all software products serve as a means of learning how to use the computer.)

The items have a rating spectrum of five degrees. Some items (e.g. Does the game have a tutorial?) can only be answered with yes or no.

4 EXPERIMENT I – CHECKLIST (STUDENTS)

4.1 Method

Subjects
Ten undergraduate psychology students - five women and five men. The average age was 21.8 years.

Material
The three edutainment games selected for the study were:
Max und das Schloßgespenst (Tivola).
Gus geht nach Cyberopolis (T1 New Media).
Zuppel und Guppi: Das Geheimnis im versunkenen Schiff Zylox (B.I.M.).

Procedure and design

Each student judged all three edutainment games on the basis of the checklist. The order in which this was done was balanced. The subjects played each game for approximately one hour. Afterwards they answered the questions on the checklist and returned to the game when necessary. The testing of one edutainment game took approximately three hours.

4.2 Analysis and results

The items were coded from 1 to 5; positive = 1, negative = 5. (The 'yes/no' questions were coded 1 and 5 respectively.)

Item selection

The item selection was carried out in two phases:

1. *ANOVA (analysis of variance)*
   For each individual item the play factor and the subject factor were checked in terms of exerting a significant impact. Ideally, the play factor has a significant influence and the subjects do not. The items for which the play factor exerted a significant influence (while the subject factor did not) remained on the checklist.
   After this analysis 89 items were filtered from the original 236 items.
2. *Item analysis*
   Items which were not selective, i.e. items with an item discrimination index of less than 0.3, were eliminated.
   Next, the difficulty of the items was analyzed. Only items having a difficulty index of between 0.2 and 0.8 were accepted.
   55 items remained.

The new checklist

The original checklist was reduced from 236 to 55 items.

The following is a description of the remaining items (the questions were originally posed in German):

- *Cover and booklet*
  Cover information.
  Booklet: clarity, thoroughness, presence of examples, explanation of educational aims, motivation value.
- *Entertainment value*
  Whether the following are offered: fun, things of interest, high number of animations, varying animations, atmosphere, background story, objects of game, long playing time, varying feedback, sufficient praise, performance checks, chance to succeed via player's achievement, quick screen display, attuning of music and sound to game actions; and what the general entertainment value is.

- *Suitability for children*
  Whether the game appeals to a child, how appropriate the tasks are for a child, how appropriate the text is for a child.
- *Ease of use*
  Whether the following are present: highlighting of icons when 'touched' or clicked on, visibility of text against background, overview of game rules, explanation of game rules, method(s) of navigation, sufficient online help, help function, help function in every situation, does the game facilitate remembering the important things, adaptability to personal needs, different levels of difficulty, method of stopping or skipping processes within the game; and what the general ease of use is.
- *Load*
  Whether the demands of the game can be fulfilled, how heavy the cognitive load is, and whether there is a balance of demands placed by the game.
- *Educational value*
  Whether the following are promoted: different abilities, independent thinking, the child's reading his or her first individual letters or words, the learning of foreign words or sentences, communication abilities, cooperation.
  Whether the tasks are: various, meaningful, of educational value.
  Whether text is: read aloud, read in a way that a child can follow the text on the screen.
  How high the educational value in all is.

Checklist results

The checklist results were based on the 55 items selected. The arithmetic means of all the items were computed. The items were not yet weighted.

The arithmetic means (low values are positive, high values are negative) were:

Max	2.36
Gus	2.04
Zuppel	3.57

'Gus' attained the best score, 'Max' the second best, and 'Zuppel' the worst. All differences between the games were significant.

Interrater reliability

The interrater reliability was 0.92 (one subject had a correlation of only 0.66 with the nine other subjects; without this subject the interrater reliability would be 0.98).

4.3 Discussion

Only those items which prompted different results for the three games remained in the final checklist version. The difficulty here is that the three selected games did not differ in all items and all dimensions. Many of the questions in which the games did not differ are very important, e.g. the question regarding violence content, as well as most of the questions in the dimension 'load'. None of the three

games exposes the user to violence or is excessively demanding. Thus the items representing these aspects were eliminated.

Owing to the importance of some of the 236 original items they should not necessarily be ignored. Items having a low selectivity or which reveal extremely high difficulty scores can also prove useful for the evaluation. With as few as three games being employed in the study, it cannot be expected that all the items succeed in defining each game separately and that all five possible answers for each item be checked with equal frequency.

The results obtained from the checklist, now containing 55 items, show significant differences between all three games. The game 'Gus' was most preferred by the adults. This is probably due to the fact that it offers the highest educational potential.

These results must be conservatively assessed because of the low number of games tested and owing to the fact that the results were influenced by the selection of the three games.

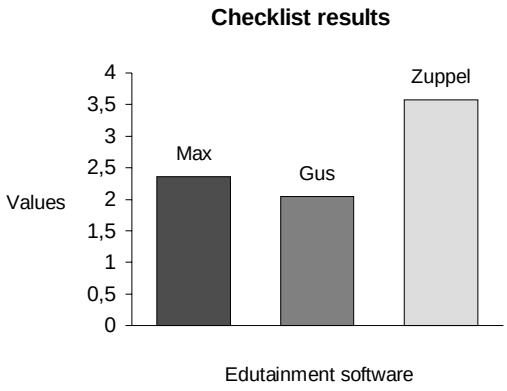

Figure 1 Checklist results (high values indicate insufficient quality, low values indicate high quality).

5 EXPERIMENT II – BEHAVIORAL DATA AND INTERVIEW (CHILDREN)

5.1 Method

Subjects

Ten pupils of the second grade took part in the study. They were between seven and eight years of age and comprised nine boys and one girl.

Material

Max und das Schloßgespenst (Tivola).
Gus geht nach Cyberopolis (T1 New Media).
Zuppel und Guppi: Das Geheimnis im versunkenen Schiff Zylox (B.I.M.).

Observation program

To simplify the study, a VisualBasic 3.0 program was written for the observation of the children. On the program surface 19 buttons with the observation variables could be seen. Whenever a subject behaved in a manner corresponding to an observation variable, the appropriate button was then clicked with the mouse. The program counted the clicking frequency and computed the quotient of positive and negative observations.

- *Positive observations*
  Laughter, smile, joy, exclaimations/sounds (neutral), comment to observer, comment to computer, comment to oneself, pride, fascination.
  Commenting shows that the child is interested in the game and is therefore to be judged positively.
- *Negative observations*
  Aggravation, anger/rage, disappointment, sigh, disquietude, question, helplessness, looking away, boredom, exhaustion.
- *Additional buttons*
  A pause button (for every kind of pause, also implemented when subject wanted to take a break).
  A button for entering text.

Interview

When the game was over the child was asked fourteen questions in connection with the following factors: fun, identification, subjective achievement, difficulty, exhaustion.

Procedure and design

The children were called in from their school lessons one by one. In a room at the school they played 'Max' and 'Gus' on a multimedia PC for about 50-60 minutes and 'Zuppel' for about 20-30 minutes. (The game 'Zuppel' requires only 20-30 minutes to play.) While playing the children were observed by the author with the help of the program. After each game they were interviewed for 15 minutes. The sequence in which the games were played, as well as the time of day (a.m. and p.m.), was balanced. When all three games were over every child was asked which game he or she liked most, second most, and least (preference judgement).

5.2 Analysis and results

Preference judgments of the children

- Seven children preferred 'Max'.
- Two children preferred 'Gus'.
- One child could not decide between 'Max' and 'Gus'.
- Eight children liked 'Zuppel' least.

Interview

The interview questions were credited with 0, 1, or 2 points. Arithmetic means were computed for every game.

Only the differences between 'Max' / 'Zuppel' and 'Gus' / 'Zuppel' were significant.

Max	1.74
Gus	1.76
Zuppel	1.52

Behavioral data

The program computed the frequencies. The variable 'look away' was not retained because the children looked away from the screen to contact the observer.

The quotient was computed from the remaining positive and negative observations. All children displayed more positive than negative behavior. The one extreme value (22.67: almost four standard deviations from the average) was excluded. Thus data from nine children were used. The behavioral data appeared to correlate with the children's preferences since 'Max' had the highest value. But statistically only the differences between 'Max' / 'Zuppel' and 'Gus' / 'Zuppel' were significant.

Max	6.08
Gus	4.48
Zuppel	2.81

5.3 Discussion

The results of the interview and the observation were rather similar. In both cases the only significant differences were between 'Max' / 'Zuppel ' and 'Gus' / 'Zuppel'. 'Max' and 'Gus' scarcely differed statistically from each other.

The results of the observation were more clear than those of the interview. Observation seems to be a very suitable method when children are test subjects.

The preference question ('Which game did you like best?') produced the clearest results and demonstrated the preference for 'Max'. Unfortunately, these results could not be used for the statistical analysis.

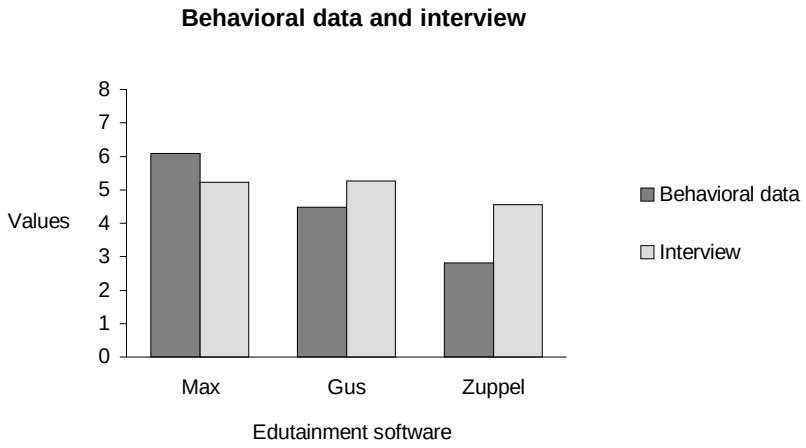

Figure 2 Behavioral data and interview (The interview values were transformed graphically so as to present them in the diagram.)

6 CORRELATION BETWEEN CHECKLIST AND CHILDREN'S DATA: REGRESSION ANALYSIS

The regression analysis shows the type of correlation existing between two variables. Thus the value of the dependent variable – the behavioral data and the interview– can be predicted by using the value of the independent variable – the checklist.

If it is possible to predict the children´s data by using the checklist results, the checklist will be validated.

Linear regression analysis

There is scarcely a correlation between behavioral data and interview results. They apparently measure two different things. Therefore, two analyses of regression were computed. The behavioral data seem to represent the true feelings of the children much more so than the interview data. The differences between the games are more prevalent in the behavioral data. Observing behavior is a generally more suitable method for young children than interviewing.

6.1 Behavioral data

The checklist results of all ten students and the behavioral data of nine children were used for the regression analysis. The arithmetic mean for each game was computed to produce three checklist values and three behavioral values.

The following equation was computed:

Behavioral data = - 1.54 * checklist results + 8.55.

$$y = - 1.54\ x + 8.55. \quad (1)$$

The checklist values are reversed the polarity of the behavioral data. A high checklist score is a negative value, a low checklist score a positive value. The straight line is therefore drawn from top left to bottom right and does not begin at the origin. The coefficient b in the equation therefore has a negative sign.

The r square shows the quality of the adaptation of the regression line. It represents the proportion between explained variance and total variance (sum of squares regression / sum of squares residual). R square is always between zero and one. In this case the adaptation was satisfying: r square = 0.58.

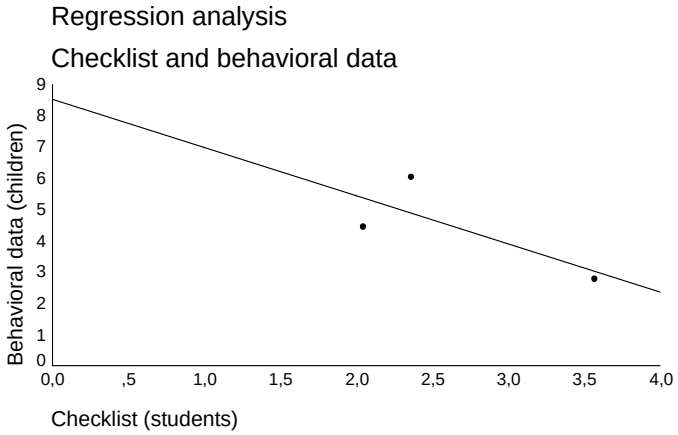

Figure 3 Regression analysis (behavioral data).

6.2 Interview

The checklist results of all ten students and the interview data of ten children were used for the regression analysis. The arithmetic mean for each game was computed to produce three checklist values and three behavioral values. In this case the adaptation was very good: r square = 0.98.

The following equation was computed:

$$y = -0.16\,x + 2.11. \quad (2)$$

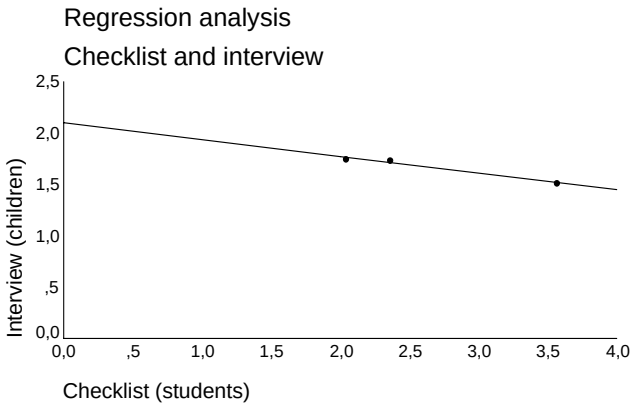

Figure 4 Regression analysis (interview).

6.3 Discussion

The results of the regression analysis must be considered based on the fact that only three games could be examined and therefore the line consists of only three points. The two different regression analyses demonstrated different qualities of adaptation. The children's interview data correspond most to the checklist results. The children's behavioral data were less successful although the adaptation was satisfactory.

Perhaps answering questions – whether in an interview or with a checklist – produces similar data while observation of behavior basically offers a completely different quality of data. One could speculate that observing the behavior of adults may provide values similar to those obtained when children are observed.

7 TEST CRITERIA

The design of the checklist is based on the principles of the classic test theory (CTT) model.

Reliability (internal consistency)
Cronbach´s alpha = 0.97.

Validity (criterion-related validity)
Correlation coefficient = 0.76 (criterion = children´s behavior).

Correlation coefficient = 0.99 (criterion = children´s interview).

8 GENERAL DISCUSSION

The results allude to a few general problems involved in constructing a checklist for adults to evaluate children's software. The preferences of the adults and those of the children differed slightly. The adults preferred 'Gus'. Some of the aspects on which the adults put emphasis were different from those placed by the children. Thus, the high educational value of the game 'Gus' seemed to be the main appeal for the adults. The children clearly preferred 'Max' when being asked directly which game they liked best (eight of the ten children). This result was also apparent in the behavioral data although there it was of no statistical significance. In the interview, no difference at all occurred between 'Max' and 'Gus'. In both children's data – the behavioral data and the interview – there were no significant differences between 'Max' and 'Gus', only between 'Max' / 'Zuppel' and 'Gus' / 'Zuppel'.

In order to standardize the checklist, ten to fifteen edutainment games and roughly twenty adults and children would be necessary. It would also be useful to have software experts rather than students for the study.

The study must be seen as a kind of exploratory study indicating a tendency. The results are connected with the selected games: the elimination of the items is related to the three games.

The regression line shows that it is generally possible to predict the children´s values with the checklist. Thus an expert would be able to judge the quality of children´s software with the help of the checklist. He or she could roughly predict how a child would react to a certain software. The checklist could also be used as a guideline when designing software for children.

However, the involvement of children in the designing and evaluating of children's software is indispensable for a well-balanced decision.

9 OUTLOOK

For a practical application the checklist procedure must again be considered. A three-step scale would perhaps be more appropriate. The items must be weighted. Furthermore, they could be summarized to different modules so that the checklist could be adapted to different kinds of software by selecting only certain modules.

A validation with a greater number of subjects and edutainment games would be desirable.

10 REFERENCES

Druin, A. (1997). The CHI97. CHIkids Program: A Partnership between Kids, Adults and Technology. *interactions*, september + october 1997, 48 – 59.

Feibel, T. (1997). *Kindersoftware-Ratgeber 1998*. München: Markt & Technik.

Fritz, J. & Fehr, W. (1996). Wie wir Computerspiele beurteilen. In Jugendamt der Stadt Köln (Hrsg.), *Computer- und Videospiele – pädagogisch beurteilt*, Band 5 (S. 10-13). Bonn: Bundeszentrale für politische Bildung.

Geisler, T. (1995). *Kids im Computer.* Berlin: BB Jugend + Computer, Förderverein für Jugend- und Sozialarbeit e.V.

Hanna, L., Risden, K. & Alexander, K. (1997). Guidelines for Usability Testing with Children. *interactions*, september + october 1997, 9-14.

Lerchenmüller-Hilse, H. (1995). *Computerspiele - Spielspaß ohne Risiko.* Köln: Arbeitsgemeinschaft Kinder- und Jugendschutz (AJS).

Lern- und Spielabenteuer auf CD-ROM – Timmy und das Löwenkind. *L.A. Multimedia - Magazin für Medien und Bildung* (1997). Heft 3, August 1997. Braunschweig: Westermann.

Oppermann, R.; Murchner, B.; Reiterer, H.; Koch, M. (1992). *Software-ergonomische Evaluation: Der Leitfaden EVADIS II.* Berlin, New York: Walter de Gruyter.

Prümper, J. & Anft, M. (1993). *ISONORM 9241/10: Beurteilung von Software auf Grundlage der Internationalen Ergonomie-Norm ISO 9241/10.*

Robertson, J. W.(1994). Usability and Children´s Software: A User-Centered Design Methodology. *Journal of Computing in Childhood Education*, 5 (3/4), 257 – 271.

Zey, R. (1994). *Bildschirmspielereien: Der Elternratgeber über Video- und Computerspiele*. Weinheim (u.a.): Beltz.

11 BIOGRAPHY

Born in Munich in 1968. Study of psychology in Regensburg and Berlin. Advanced studies in engineering psychology and minor in computer science. Special interests: software for children, software ergonomics, the Internet, computer science.