Interactive Learning Environments
Vol. 20, No. 6, December 2012, 501–512

Effective electronic materials: are teachers aware of these?

P. Luik*

Department of Education, University of Tartu, Tartu, Estonia

(*Received 9 February 2010; final version received 27 August 2010*)

This study analyses to what extent teachers recognise which interactive multimedia software is efficient and which is not. The results are based on two correlation studies. The first study was carried out with 35 different pieces of interactive multimedia software for secondary students, and 34 pieces of interactive multimedia software for primary students were used in the second study. The teachers of special subject and class-teachers participated in these studies. The correlations between the teachers' evaluations about the interactive multimedia software and students' learning outcomes were found. The present study revealed that primary teachers' evaluations were not related to or negatively correlated with students' learning outcomes. Teachers of special subjects were more able to recognise the efficiency of educational software, but not in the case of all features and all groups of students.

Keywords: interactive multimedia; teachers' education; experiment; primary education; secondary education

1. Introduction

In Estonian schools, the teacher chooses the learning materials for students. Whether the students use effective or ineffective learning materials, including multimedia software, and whether they use ICT in the learning process at all depend on the teacher. Besides the personal characteristics of the teacher (conservative or innovative, ready to take risks, etc.), the teacher's expectations about the particular software influence his/her computer skills and attitudes towards the usage of computers and multimedia software in the learning process. The teachers' attitude towards the usage of computers in learning may be as positive as it can be; the teachers may have good computer-skills, but the multimedia software is not used by him/her if he/she believes that particular software is inefficient and students cannot achieve high learning outcomes by using such learning materials.

Traditional learning materials (text-books and work-books on paper) are composed by competent authors in Estonia. All traditional learning materials are evaluated by experts, and these materials should get approval from the Ministry of Education. With the implementation of computers in schools, the number of different multimedia software has rapidly increased in Estonia. Multimedia software,

*Email: piret.luik@ut.ee

ISSN 1049-4820 print/ISSN 1744-5191 online

http://dx.doi.org/10.1080/10494820.2010.533684
http://www.tandfonline.com

financed by the Tiger Leap Foundation, is also evaluated by the experts. But there are learning portals (such as Koolielu, Miksike, LeMill), where teachers can upload digital multimedia learning materials that they have composed and download the materials composed by the other practitioners. This kind of learning material is not evaluated by the experts. Moreover, none of the pieces of the multimedia software is approved by the Ministry of Education. Our teachers have experience with the traditional learning materials, and they also were taught using such materials when they were in school themselves. Squires and Preece (1999) have claimed that teachers need to be able to evaluate predictively educational software, too. However, the question arises: are our teachers as competent in choosing interactive multimedia materials? Do they know which interactive multimedia software is effective and which hinders learning? The study was carried out to answer these questions and to find the relationships between the teachers' evaluations and students' learning results.

2. Review of literature

A greater variety of techniques for evaluating software has been developed to evaluate commercial software (Squires & Preece, 1999). The first major effort to evaluate educational software was carried out by *Educational Products Information Exchange* (EPIE). EPIE is a non-business organisation that has inspected textbooks, audio–video materials and other educational resources since 1967. In the 1980s, they started to oversee educational software, and records about these results were retained in the database TESS (Buckleitner, 1999).

Papers about the evaluation of the educational software were available in the press as well. In 1983 in the journal *The Computing Teacher*, Singer's paper '*How do teacher and student evaluations of CAI software compare?*' was published. In this article the author recommends testing educational software in schools and disseminating the testing results with particular software. However, this kind of research is not widely conducted (Higgins, 2000).

The first standardised evaluation instruments for educational software appeared in the 1980s. These were *High/Scope Educational Research Foundation* and Haugland's and Shade DAP Scale. These kinds of instruments intended to quantify factors of software design that relate to the effectiveness of products for young children (Buckleitner, 1999). Such scholarship from the 1980s to the present suggested that the evaluation of the educational software and the research about the evaluation of the educational software are important.

Mostly the evaluating techniques are for predictive evaluation, which occurs when teachers are either planning lessons or making purchasing decisions (Squires & Preece, 1996). These kinds of evaluating techniques use mostly checklists (Tergan, 1998) and some of them use rubrics (Squires & Preece, 1996). Squires and Preece (1996) claimed that when using these checklists or rubrics teachers need to be aware of learning and usability and their integration, but they know little about the usability. Ngu and Rethinasamy (2006) have declared that it is inadequate to use only teachers' evaluations because they include only teachers' views of the software. It is important to invite students to participate in the evaluation process. When learners are participating in the process of evaluating educational software, evaluators are able to find out whether the learners acquire the skills a software program is designed to teach.

There is some research that deals with teachers' evaluations (e.g. Bos, 2003; Kurz & Middleton, 2006; Lavonen, Aksela, Juuti, & Meisalo, 2003; McCormick & Li, 2006), students' evaluations (e.g. Diederen, Gruppen, Hartog, & Voragen, 2005; Herring, Notar, & Wilson, 2005) and both students' and teachers' evaluations (e.g. Muwanga-Zake, 2007; Roth, Fagan, Griffith, Nelson, & Zhao, 2003). Nurmi and Lehti (2003) have investigated a little over 500 European teachers' expectations about the usefulness of digital learning materials. Darby (2007) has investigated students' evaluations on courses and found that halo effects occur, which means that students who reportedly like one aspect of a course also appear to like another, and those who reportedly dislike one aspect also appear to dislike others. Some studies deal with students' learning outcomes achieved by working with these learning materials and their evaluations (e.g. Jacobson, 2006; Ngu & Rethinasamy, 2006). But although Ngu and Rethinasamy (2006) collected both tests and questionnaires of students, they did not find any relationships between these data and used the data only for evaluating the educational software. Bangert-Drowns and Pyke (2002) conducted a study in which they correlated teachers' ratings with student reading test scores. Based on direct observation, the two 5th-grade classroom teachers and the computer teacher rated how frequently each student engaged in seven different kinds of interaction with software. The researchers found a significant correlation ($r = 0.62$, $p < 0.01$) between the classroom teacher's ratings of student engagement and students' standardised reading test scores.

The broadest evaluation was described by Stoller, Horn, Grabe, and Robinson (2006). This collective of researchers collected triangulate data from three participant groups: pilot faculty, pilot students and potential adopters, who served as external reviewers. The surveys, interviews, assessments, research papers and written reviews were used for an interdisciplinary textbook development project. The goal of this research was to improve and develop the project, not to find whether these evaluations are in accordance with each other or not.

Some decisions and developmental projects are grounded on teachers' evaluations as well. For example, in some cases, educators used teachers' evaluations for developing new software products (e.g. Lavonen et al., 2003). The other example is the work of Herring et al. (2005), who developed an evaluation form for teachers using evaluations of pre-service teachers. Herring et al. (2005) used students' evaluations for composing the evaluation form for teachers. But designs of these kinds of studies do not prove that teachers' evaluations are reliable and could be taken as a ground of the project or study.

It could be concluded that there are few studies that deal with the relationships between the teachers' evaluations of the learning materials and students' learning results achieved by working with these learning materials. It is an important area, because if the teacher composes or chooses ineffective learning material, believing that it is efficient, the students cannot achieve a high learning outcome. Without evidence that working with software that is getting high teachers' evaluations, students' learning outcome increases, there is no reason for using teachers' evaluations as the basis for developing new products. 'The evaluation of educational software has become a critical information skill for teachers' (Herring et al., 2005, p. 100). Kurz and Middleton (2006) wrote that software evaluation should focus on student learning, aligning with the way students learn along with usability.

Informal predictive evaluations rely on past personal experience (Squires & Preece, 1999), but the question arises, are these kinds of evaluations, which are based on personal experience of teachers, reliable? Therefore, the *objective* of this study was to reveal the link between teachers' evaluations and students' learning outcomes working with the educational software. Because in Estonia, class-teachers spend more time with their students than teachers of special subjects and for that reason they should know their students better, *the hypothesis of the study* was: Class teachers' evaluations have more positive correlations with their students' learning outcome than evaluations of teachers of special subjects.

3. Method

3.1. Samples

The results are based on two studies. The first study was carried out with 35 different pieces of text-based educational software – electronic textbooks. Thirty-four pieces of drill-programs were used in the second study. One piece of educational software was considered as the unit in the particular educational software, which was acquirable approximately during a half of the lesson time. In the study, educational softwares about different subjects were used. There were 20 pieces of educational software about mathematics, 19 pieces about English as a second language, 12 pieces about history, 6 pieces about geography, 6 pieces about chemistry and 6 pieces about Estonian together in two studies.

Participants of the first study were 10th form students (age 15–16) from four schools in Estonia. Fifty-four students (21 boys and 33 girls) participated in this study. The second study was carried out with students from the 3rd form (age 9–10) in four schools. Eighty students (37 boys and 43 girls) participated in this study. Both student groups were of mixed ability. The high-achieving and low-achieving students were determined according to the results of the tests. There were 19 low-achieving and 19 high-achieving students in the sample of the first study and 18 low-achieving and 31 high-achieving students in the sample of the second study.

The teachers of special subjects and class-teachers also participated in these studies. In the first study, the history teacher evaluated the educational software about history, the Estonian teacher evaluated the educational software about Estonian, and so on. In the second study, the class-teacher evaluated the drill-programs. Each piece of educational software was evaluated by 20 teachers. Amongst the teachers of special subjects there were both sexes, women (77%) and men (23%), but all class-teachers were women. The age of participating teachers varied from 27 to 53 years ($M = 39.2$, SD = 7.4).

These students and teachers were experienced in learning with computers. Only 20% of the 10th form students mentioned before the study that their computer skills were not good; not a single one of them evaluated his/her skills very bad. Only 14% of the 3rd form students mentioned that their computer skills were not good before the study. Participating teachers were selected using educational technologists' evaluations and their own evaluations of their computer skills. Participating teachers and educational technologists of their schools got the checklist, where they had to evaluate on 5-point scale (5 – very good .... 1 – missing) different computer skills and knowledge: searching information, communicating, composing worksheets, composing presentation, knowing computer terminology, typing, understanding computer messages, ability to apply computer-based learning methods, technical skills. All participating teachers and their educational technologist evaluated all their computer skills and knowledge at least as fair (3 points on 5-point scale). All the participating teachers had graduated from at least one course related to the usage of computers in classroom. All courses provided knowledge about how to use basic computer applications. All the teachers in these studies had used computers in their classroom activities. Sixty per cent of teachers of special subjects and all class-teachers had used computer applications for preparing digital learning materials for their students as well.

3.2. Instruments

Puustinen, Baker, and Lund (2006) claimed that there can be several reasons for evaluating educational multimedia by teachers:

(1) to decide which is most adapted for usage in teaching a particular subject in a particular class;
(2) to give feedback about the effectiveness of educational software during the process of its development;
(3) to redesign educational multimedia.

The authors (Puustinen et al., 2006) stated that for different reasons different tools are needed. For example, for evaluating in order to decide which is most adapted for usage in teaching a particular subject in a particular class, a checklist of important criteria could be the best tool, but for evaluating effectiveness during the process of its development, an experiment could be used. In this study, the main reason for evaluating educational multimedia by teachers was to decide which is most adapted for usage in teaching a particular subject in a particular class by the opinion of teachers. Therefore, the checklist for the teachers was used. The teachers evaluated, on a 10-point scale, the effectiveness of the particular educational software, intelligibility of the content, plainness of the structure, pleasantness and interest of the learning material, possibilities of motivating the students and simplicity of manipulating and appropriateness of feedback.

Ngu and Rethinasamy (2006) observed that during the evaluation of the software, it is good if students are pre-tested, observed as they work through the software program, and post-tested. The authors gave reasons for this kind of study design: if students serve as participants in the evaluation process, other evaluators (e.g. teachers) observe the students as they use the software, test what students have learned as a result of using the software, and draw conclusions about the quality of the software. Therefore, pre- and post-test for students were used in these studies as well.

In these studies, the tests were composed by experienced teachers (not the same ones who participated in the studies). The tests were in two versions, and both forms of the test were on paper. As the tests differed in the number of items, the percentage of the students' score was calculated. The *reliability* (Cronbach's α) of the tests was 0.76–0.86 and the *validity* was evaluated by the experts.

In the first study were factual questions and questions that required analysis or argumentation or transfer of information in a new situation. These were program tests where all answers could be found from the texts of electronic textbooks and all subtopics of the particular unit were covered by the questions. There were different kinds of questions in all the tests in the first study: multiple-choice, free-response and matching questions. For some of the questions, there were several right answers.

The questions about the basic skills (multiplication tables or adding or subtracting, translating the words from English to Estonian and *vice versa*) were included in the tests of the second study. Recalling information and recognition of information were needed in both tests. The tests for third graders were paper analogues of the drill questions and the test questions for the tenth graders were similar to the self-assessment questions of the electronic textbooks. The paper-and-pencil tests were used because the meta-analysis (Wang, Jiao, Young, Brooks, & Olson, 2007, 2008), indicates that the difference between students' reading achievement scores and mathematics scores from computer-based test and paper-and-pencil administrated tests was not statistically significant.

3.3. *Procedure of the experiments*

Students filled in pre-tests before studying the unit. With the pre-tests, the students' prior knowledge was measured. After that, the students worked with the pieces of educational software independently. After studying the unit, they filled in the post-tests; results of the tests were collected as evidence of learning achievement. All of the students in one study worked with the same units, and all the students were asked to study all pieces of the educational software used in particular study. So all the students of the first experiment studied 35 pieces of the electronic textbooks, and the students of the second experiment studied 34 pieces of the drill-programs. At the same time, the teachers gave rankings for the different aspects of the same educational software.

3.4. *Data analysis*

The statistical package SPSS 11.5 for Windows was used for data analysis. The mean test score of all the students and the mean evaluations of the teachers were calculated in the case of each piece of educational software. Also, the mean test score of boys and girls, high- and low-achieving students, was calculated. Besides the teachers' evaluations on different aspects of the educational software, the total evaluation was used. The total evaluation was calculated as the mean value of the different aspects of evaluations (evaluations of the effectiveness of the particular learning program, intelligibility of the content and plainness of the structure, pleasantness and interest of the learning material, possibilities of motivating the students, simplicity of manipulating and appropriateness of feedback).

Data were analysed using correlation analysis in order to test if there were significant relationships between the different aspects of evaluations of teachers and the test scores of students. Also, the correlation coefficients between the teachers' mean evaluations and the mean test scores of boys and girls, high- and low-achieving students, were calculated.

4. Results

The Pearson correlation coefficients between the different aspects of mean evaluations of the teachers of the 10th grade and students' test scores are given in Table 1, and the Pearson correlation coefficients between the different aspects of mean evaluations of the teachers of the 3rd grade and students' test scores are given in Table 2.

Table 1. The Pearson correlation coefficients between the different aspects of mean evaluations of the teachers of the 10th grade and 10th grade students' test scores.

Aspect of the evaluation of the teachers of 10th grade	Correlation coefficients with the test scores of all students	Correlation coefficients with the test scores of boys	Correlation coefficients with the test scores of girls	Correlation coefficients with the test scores of high-achieving students	Correlation coefficients with the test scores of low-achieving students
Effectiveness of the learning program	0.45**	0.48**	0.40*	0.46**	0.38*
Intelligibility of the content	0.21	0.19	0.20	0.41*	0.06
Plainness of the structure	0.47**	0.47**	0.43**	0.52**	0.26
Pleasantness of the learning material	−0.08	0.03	−0.14	0.17	−0.22
Possibilities of motivating the students	−0.15	−0.01	−0.23	.29	−0.47*
Interest of the learning material	−0.34	−0.16	−0.36*	0.12	−0.52**
Simplicity of manipulating	0.12	0.08	0.13	0.25	−0.01
Feedback	−0.40*	−0.14	−0.53**	−0.11	−0.45**
The total evaluation	−0.02	0.12	−0.10	0.40*	−0.26

Note: *Statistically significant at the 0.05 level. **Statistically significant at the 0.01 level.

5. Discussion

As the results indicate, there are a few statistically significant correlation coefficients in Table 1 (17 significant relationships from 45) and even fewer in Table 2 (5 significant relationships from 45). Unfortunately, all 3rd grade teachers' evaluations, which were significantly related to the students' learning results, were negatively correlated. This finding means that when the teacher has the opinion that the particular educational software is good for some aspect of learning, the students achieve poorer results of learning when working with this drill-program, and the opposite: when the teacher believes that the particular educational software is inadequate for some aspect of learning, the students gain better results of learning when working with this drill-program.

Despite the fact that class-teachers spend more time with their students than teachers of special subjects, and for that reason should know their students and should be able to choose suitable educational software for their students (Mei-Yan, Walker, & Huang, 1999), these experiments indicated the opposite result. The 10th grade teachers' evaluations of the effectiveness of the electronic textbooks were related to the learning outcomes of students, but the class-teachers' (3rd grade) evaluations of the effectiveness of the drill-programs were not. If the teachers of special subjects evaluated the electronic textbook higher, the students got higher learning results. But there is no reason to believe that if the

Table 2. The Pearson correlation coefficients between the different aspects of mean evaluations of the teachers of the 3rd grade and 3rd grade students' test scores.

Aspect of the evaluation of the teachers of the 3rd grade	Correlation coefficients with the test scores of all students	Correlation coefficients with the test scores of boys	Correlation coefficients with the test scores of girls	Correlation coefficients with the test scores of high-achieving students	Correlation coefficients with the test scores of low-achieving students
Effectiveness of the learning program	−0.04	−0.17	0.05	−0.02	0.10
Intelligibility of the content	0.15	0.02	0.21	0.09	0.26
Plainness of the structure	0.21	0.20	0.18	0.30	0.35
Pleasantness of the learning material	−0.07	−0.02	0.09	−0.20	−0.02
Possibilities of motivating the students	−0.49*	−0.36*	−0.51**	−0.41*	−0.19
Interest of the learning material	−0.32	−0.40*	−0.24	−0.20	−0.17
Simplicity of manipulating	−0.00	−0.03	0.01	0.22	0.12
Feedback	−0.05	0.04	−0.10	−0.26	−0.11
The total evaluation	−0.33	−0.31	−0.31	−0.33	−0.15

Note: *Statistically significant at the 0.05 level. **Statistically significant at the 0.01 level.

class-teachers evaluated the learning program higher, the students would get higher results. Or if the class-teacher considers the particular drill-program to be ineffective for learning, the students cannot get good results working with this learning program. One explanation of such results could be the different type of the educational software as well. Some factors related to the presentation of items, such as computer interface, item layout, and graphics in computer-based tests may result in differences in student's performances between computer-based test and paper-and-pencil tests (Wang et al., 2008) and class-teachers decided how effective the drill is according to the students' performances of the drill sessions. The self-assessment of the electronic textbooks was more similar to the paper-and-pencil tests than drill sessions and therefore the 10th grade teachers were able to more precisely evaluate the effectiveness of the electronic textbooks.

Also, the 10th grade teachers' evaluations about the plainness of the structure of the material were positively related to the learning results of all students: boys, girls and high-achieving students. The teachers of special subjects were able to identify which content was plain for students, except the low-achieving students. This result is in accordance with the statement of Tergan (1998), who wrote that all learners do not have the same competences and therefore some items on evaluation checklists are suitable only for some groups of students. But the class-teachers' evaluations about the plainness of the structure of the drill-program were not related to any test scores of the students. Therefore, the class-teachers did not recognise the plain structure of the drill-program for the 3rd grade students.

One of the aspects of the evaluations in this study was the intelligibility of the content. The structure of the educational software could be plain, but not intelligible for the students. If the 10th grade teachers evaluated the intelligibility of the electronic textbooks higher, only the high-achieving students got higher test scores. Also, the total evaluation of the teachers of special subjects was positively correlated only with the high-achieving students' test score. Class-teachers' evaluations about the intelligibility of the learning program and total evaluation were not related to any test score of students. Explanation of this result could be also the statement of Tergan (1998) that teachers could not evaluate such kind of aspects of educational software taking into account competences of all students. Therefore, it could happen that if the teachers tried to evaluate considering different students' needs, the ranking did not relate to any students group.

It is interesting, that if the teachers of the 10th grade evaluated the educational software as motivating for the students, the low-achieving students' test scores were lower. But if the 3rd grade teachers evaluated particular learning programs as motivating the students, the test scores of all students, boys, girls and high-achieving students, were lower. The reason for that result might be the fact that 3rd grade teachers evaluated the possibilities of motivating the students more highly when the learning program was a game and was more attractive. The teachers did not discern the risks of the attractiveness. Maybe the teachers saw that children were better motivated and their enjoyment was higher when using drills which are more like games and more attractive than the other drills, but the high level of enjoyment did not assure high learning outcomes. Motivation is a necessary, but not sufficient, condition to learning and there are few positive results of gaming, at least as compared to traditional classroom instruction (Rieber & Noah, 2008). To the contrary, the 10th grade teachers evaluated the possibilities of motivating the students more highly when the electronic textbook offered more information.

Also, the evaluations of the interest of the learning material were negatively related to the students' test scores. The 10th grade teachers did not anticipate the interest level of the girls and low-achieving students (there were significant negative correlations between the teachers' evaluations and the girls' and low-achieving students' test scores). They did not know which material and which design of learning programs was interesting for boys as well (there was significant negative correlation between the teachers' evaluations and the boys' test scores). Reber, Hetland, Chen, Norman, and Kobbeltvedt (2009) declare that teachers rarely know the individual interests of their students, too.

Feedback by the computer is also different from the traditional feedback. The class-teachers' evaluations about the feedback were not significantly related to the students' test scores. But the highest number of significant negative correlation coefficients between the 10th grade teachers' evaluations and students' test scores were in this field. If the teachers of special subjects evaluated the fitness of the feedback of the educational software more highly, all students, girls and low-achieving students, achieved lower results of learning. Teachers did not recognise the effective characteristics of the computer-based feedback. They considered that the feedback should be the same as in the traditional classroom.

The teachers' evaluations about the pleasantness of the learning material and about the simplicity of manipulating with the educational software did not have any relationships with the students' test scores. The teachers did not know which learning material was pleasant for students, and they did not know the real computer skills of students.

6. Conclusion

The present study revealed that teachers were not able to recognise efficient educational software. An interesting result was that the evaluations of the teachers of the 10th grade about the effectiveness of the particular educational software were significantly and positively correlated to the students' test scores, but the evaluations of class teachers (in the 3rd grade) were negatively related to the students' learning results. So, the hypothesis was not proved.

The 10th grade teachers' evaluations were more connected with the boys' and high-achieving students' learning outcomes, but there was not such a trend among the class-teachers' evaluations. An explanation for that result might be the differences due to the different grades of the students. In first grades, it is essential that every student cope with the learning. On the other hand, the attention is more on the high-achieving students in the upper grades of Estonian schools.

The teachers were educated in the field of computer literacy, but they did not precisely discern the efficiency of educational software. There is a reason to suppose that they are also not able to compose efficient digital learning materials (PowerPoint presentations, educational web-sites, etc.) themselves as well. Teachers need skills to discriminate between the constraints and affordances of educational software along with the ability to use various applications to achieve certain goals. Clyde (1998) and Herring et al. (2005) point out, that besides the ability to select material for usage in a particular curriculum unit, teachers have to teach students how to evaluate electronic materials. Evaluators should take into consideration cognitive preconditions of the learner and check whether learners with particular cognitive preconditions would find the content plain or easy to understand (Tergan, 1998). Teachers must have the skills to critically evaluate the content, presentation of the information, and different features of interactive multimedia software for different students. Therefore, the courses for pre-service and in-service teachers should prepare teachers to recognise efficient educational software, to compose efficient digital learning materials, and to teach evaluation criteria for electronic materials to students, and these courses are needed in teacher education.

In the present study, correlation analysis was used. The weakness of correlational analysis is that correlations do not necessarily imply cause–effect relationships. This study tested if the evaluations of teachers, which rely on personal experience, are related to the students' learning outcomes, but as Squires and Preece (1999) found, mostly the predictive evaluations are used. It might be that in this study students' behaviour, questions or conversation using particular software influenced teachers' evaluations. Therefore, further studies in this area are needed to find out the relationships between the predictive evaluations and students' outcomes.

Notes on contributor

Piret Luik is an assistant professor in the University of Tartu, Estonia. Her research areas are: use and effectiveness of different forms of educational technology and e-learning; teachers' reflection, beliefs and motivation; internet risks and virtual worlds' influence on people.

References

Bangert-Drowns, R.L., & Pyke, C. (2002). Teacher ratings of student engagement with educational software: An exploratory study. *Educational Technology Research & Development*, *50*(2), 23–38.

Bos, M. (2003). Experienced teachers' expectations about the potential effectiveness of spelling exercises. *Annals of Dyslexia*, *53*, 104–127.

Buckleitner, W. (1999). The state of children's software evaluation – Yesterday, today and in the 21st century. *Information Technology in Childhood Education*, 211–220. Retrieved February 3, 2009, from http://childrenssoftware.com/evaluation.html

Clyde, A. (1998). Evaluating and selecting internet resources. *Emergency Librarian*, *25*(4), 32–34.

Darby, J.A. (2007). Are course evaluations subject to a halo effect? *Research in Education*, *77*, 46–55.

Diederen, J., Gruppen, H., Hartog, R., & Voragen, A.G.J. (2005). Design and evaluation of digital learning material to support acquisition of quantitative problem-solving skills within food chemistry. *Journal of Science Education and Technology*, *14*(5/6), 495–507.

Herring, D.F., Notar, C.E., & Wilson, J.D. (2005). Multimedia software evaluation form for teachers. *Education*, *126*(1), 100–111.

Higgins, K. (2000). Evaluating educational software for special education. *Intervention in School & Clinic*, *36*(2), 109–115.

Jacobson, E. (2006). Computer homework effectiveness in developmental mathematics. *Journal of Developmental Education*, *29*(3), 2–8.

Kurz, T.L., & Middleton, J.A. (2006). Using a functional approach to change preservice teachers' understanding of mathematics software. *Journal of Research on Technology in Education*, *39*(1), 45–65.

Lavonen, J., Aksela, M., Juuti, K., & Meisalo, V. (2003). Designing user-friendly data logging for chemical education through factor analysis of teacher evaluations. *International Journal of Science Education*, *25*(12), 1471–1487.

McCormick, R., & Li, N. (2006). An evaluation of European learning objects in use. *Learning, Media and Technology*, *31*(3), 213–231.

Mei-Yan, L., Walker, D.F., & Huang, J. (1999). Do they look at educational multimedia differently than we do? A study of software evaluation in Taiwan and the United States. *International Journal of Instructional Media*, *26*(1), 31–42.

Muwanga-Zake, J.W.F. (2007). Evaluation of an educational computer programme as a change agent in science classrooms. *Journal of Science Education & Technology*, *16*(6), 473–490.

Ngu, B.H., & Rethinasamy, S. (2006). Evaluating a CALL software on the learning of English prepositions. *Computers & Education*, *47*(1), 41–55.

Nurmi, S., & Lehti, S. (2003). *European teachers' expectations and opinions about learning objects*. Paper presented at the symposium on designing virtual learning materials, EARLI conference, in Padova, Italy.

Puustinen, M., Baker, M., & Lund, K. (2006). GESTALT: A framework for redesign of educational software. *Journal of Computer Assisted Learning*, *22*, 34–46.

Reber, R., Hetland, H., Chen, W., Norman, E., & Kobbeltvedt, T. (2009). Effects of example choice on interest, control, and learning. *Journal of the Learning Sciences*, *18*(4), 509–548.

Rieber, L.P., & Noah, D. (2008). Games, simulations, and visual metaphors in education: Antagonism between enjoyment and learning. *Educational Media International*, *45*(2), 77–92.

Roth, C.S., Fagan, M.J., Griffith, J.M., Nelson, D., & Zhao, Y. (2003). Evaluation of a worksheet to structure teaching and learning outpatient internal medicine. *Medical Teacher*, *25*(3), 296–301.

Squires, D., & Preece, J. (1996). Usability and learning: Evaluating the potential of educational software. *Computers and Education*, *27*(1), 15–22.

Squires, D., & Preece, J. (1999). Predicting quality in educational software: Evaluating for learning, usability, and the synergy between them. *Interacting with Computers, The Interdisciplinary Journal of Human–Computer Interaction*, *11*(5), 467–483.

Stoller, F.L., Horn, B., Grabe, W., & Robinson, M.S. (2006). Evaluative review in materials development. *Journal of English for Academic Purposes*, *5*, 174–192.

Tergan, S.-O. (1998). Checklists for the evaluation of educational software: Critical review and prospects. *Innovations in Education and Teaching International*, *35*(1), 9–20.

Wang, S., Jiao, H., Young, M.J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K–12 mathematics tests. *Educational and Psychological Measurement*, *67*(2), 219–238.

Wang, S., Jiao, H., Young, M.J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments. A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, *68*(1), 5–24.