The Effects of Images on Multiple-choice Questions in Computer-based Formative Assessment

Juan-Fernando Martín-SanJosé

jumartin@dsic.upv.es

Universitat Politècnica de València, Spain

M.-Carmen Juan

mcarmen@dsic.upv.es

Universitat Politècnica de València, Spain

Roberto Vivó

rvivo@upv.es

Universitat Politècnica de València, Spain

Francisco Abad

fjabad@dsic.upv.es

Universitat Politècnica de València, Spain

Abstract

Current learning and assessment are evolving into digital systems that can be used, stored, and processed online. In this paper, three different types of questionnaires for assessment are presented. All the questionnaires were filled out online on a web-based format. A study was carried out to determine whether the use of images related to each question in the questionnaires affected the selection of the correct answer. Three questionnaires were used: two questionnaires with images (images used during learning and images not use during learning) and another questionnaire with no images, text-only. Ninety-four children between seven and eight years old participated in the study. The comparison of the scores obtained on the pre-test and on the post-test indicates that the children increased their knowledge after the training, which demonstrates that the learning method is effective. When the post-test scores for the three types of questionnaires were compared, statistically significant differences were found in favour of the two questionnaires with images versus the text-only questionnaire. No statistically significant differences were found between the two types of questionnaires with images. Therefore, to a great extent, the use of images in the questionnaires helps students to select the correct answer. Since this encourages students, adding images to the questionnaires could be a good strategy for formative assessment.

Key words

Evaluation methodologies; Interactive Learning Environments; Media in Education; Improving Classroom Teaching; Multimedia Systems

I. Introduction

The main role of a teacher is to guide students during their learning process. Another of the teachers' tasks is to determine if the students have acquired the defined learning goals. Students should demonstrate that they have acquired these defined learning goals. Rating students has been a research topic for more than 70 years. The design, development, use, and interpretation of student assessment is one of the important topics in evaluation research (Arreola, 1995). Nevertheless, teaching, learning, and assessment are changing. Teaching and learning are no longer restricted to traditional classrooms (Wang et al., 2007). New learning methods are continually being incorporated (e.g. e-Learning). E-Learning refers to the use of electronic devices for learning, including the delivery of content via electronic media such as Internet, interactive TV, etc. E-Learning presents the intersection between the world of information and communication technology and the world of education (Stankov et al., 2004), or even a virtual world (Monahan et al., 2008).

Assessment can be defined as "the measurement of the learner's achievement and progress in a learning process" (Keeves, 1994; Reeves & Hedberg, 2009). The assessment of students is a core component for effective learning (Bransford et al., 2000). There are two main forms of assessment: summative and formative (Challis, 2005). Summative assessment measures what students have learned at the end of a course or after some defined period (Hargreaves, 2008). It can also refer to checking whether or not the students have met the desired learning goals or whether they have achieved the required levels of competence (Challis, 2005). Summative assessment usually includes scoring for validation or accreditation purposes. Formative assessment is applied as a source of continuous feedback to improve teaching and learning (Hargreaves, 2008). Formative assessment can also be seen as assessment for learning that takes place during instruction in order to support learning (Oosterhof et al., 2008; Vonderwell et al., 2007). Formative assessment activities are intrinsic parts of instruction that allow learning to be controlled and the instruction to be modified until the desired learning goals have been achieved (Gikandi et al., 2011). Hattie and Timperly (2007) and Nicol and Macfarlane-Dick (2006) stated that feedback is most effective when it is directly related to clearly defined learning goals, and that effective formative feedback is not only based on monitoring the progress towards those goals but that it must also encourage students to develop effective learning strategies.

Assessment can take advantage of the use of computers and internet. One of the most common computer-based assessments (CBA) is performed online; it consists of a web site where the students can reach the survey system and log in. Once they are in, they can select their answers from multiple items and they can write down open-ended questions in text boxes. When they have submitted their answers, they can also obtain a document with a statement of accomplishment about the evaluation made (Dommeyer et al., 2004). Some of the benefits of CBA are that evaluations of this kind eliminate paper costs, can be faster and easier to complete, allow efficient processing of data and are less vulnerable to influence by the faculty (Dommeyer et al., 2002b). Additionally, CBA allows adaptive testing based on the responses, which is not possible with paper-based assessments (Brown et al., 2008). Nevertheless, online assessment also has some disadvantages, such as requiring students to have technical access and to know their log-in information. Some of the students may also experience technical problems when accessing the evaluation (Anderson et al., 2005).

In our work, we have focused on online formative assessment and multiple-choice questions. In this type of questionnaire, there is usually a question and several possible answers in which the student must select only one answer. It is very common for the answers to be just text. However, images could also be used. In this paper, we have carried out a study to determine if an added image that represents/defines an object helps the children to choose

the correct answer. For the learning process, we used a computer game where the children learned about the different historical ages (Martín-SanJosé et al., 2014a, 2014b, 2015). The primary hypothesis was that there would be significant differences between using only a text-only questionnaire and a questionnaire that, apart from the text also includes images. The secondary hypothesis was that there would be significant differences between a questionnaire with images used during the learning process and a questionnaire with images that represent the item but that were not used during the learning process.

The paper is organized as follows. Section 2 focuses on the state of the art. Section 3 presents the learning method used and the tool utilized for the development of the questionnaires. Section 4 details the study. Section 5 presents the results. Finally, Section 6 presents a number of conclusions and identifies areas for future research.

II. State of the art

Computer-based assessment is not new. Two of the first systems to support assessment were PLATO (Programmed Logic for Automatic Teaching Operations) and TICCIT (Time-shared Interactive Computer-Controlled Information Television) (Rota, 1981). From there, different tools for assessment such as the following have already been presented: 1) MarkTool (Heinrich & Lawn, 2004), which allows teachers to annotate PDF documents sent by students with formative feedback (annotations can be textual and graphical; 2) EAT (Electronic Assessment System) (Rashad et al., 2008), which allows teachers to modify the content taking into account the student's answers, answer-time, and student feedback; 3) A flexible e-assessment system designed by Dube and Ma (2010), which can be adapted to different learning styles; 4) GPAM-WATA, a Web-based dynamic assessment system, in which teachers can provide students with teaching assistance (Wang, 2010); 5) FAML (Formative Assessment-based Mobile Learning) designed by Hwang & Chang (2011), which is a mobile system for local cultural learning that runs on mobile devices (PDAs). Apart from tools, other initiatives have also been presented. In 2009, a consortium of Cisco, Intel, and Microsoft launched Transforming Education: Assessing and Teaching 21st Century Skills (Cisco et al., 2009) with the goal of mobilizing international educational, political, and business communities with regard to the needs and opportunities for transforming educational assessment and instructional practices. The JISC (Joint Information Systems Committee) published an overview of technologies, policies, and practices with e-assessment in further and higher education. The JISC is also undertaking efforts to standardize assessment. Along similar lines, the IMS Global Learning Consortium presented the IMS Question and Test Interoperability Specification (IMS Global Learning Consortium, 2008).

With regard to the preference of completing online evaluations over paper ones, there is no unanimity about this preference. Several works have indicated that students prefer completing online evaluations to paper ones (Layne et al., 1999; Dommeyer et al., 2004; Anderson et al., 2005). In the study carried out by Anderson et al. (2005) when asked about their preferred evaluation format (online or traditional), over 90% of the students selected Agree or Strongly Agree in favour of the online format. Other studies contradict these data and mention that students prefer pen and paper exams to computer-based options (Llamas-Nistal et al., 2011, 2013). Other studies have argued that online evaluations tend to produce more written comments than traditional, in-class evaluations (Dommeyer et al., 2002a), and allow students to perform the evaluation collaboratively (Conejo et al., 2013), or even perform self-evaluation (Gathy et al., 1991) or personalized assessments based on their own knowledge and objectives (Lazarinis et al., 2010). Sorenson & Johnson (2003) determined that students give more and longer answers when they are performing an online assessment than when they are using a traditional paper-pencil system. Another study stated that the

online tool was easy to use, students appreciated the anonymity of the online assessment, and that evaluations of this kind allowed students to offer more thoughtful remarks than performing the traditional evaluation (Ravelli, 2000). Blended approaches have also been taken into account in previous studies. Llamas-Nistal et al. (2013) combined the benefits of the digital world with the convenience of traditional evaluation and assessment sessions. This tool may also be seen as a cost-effective alternative to computer-supported e-assessment in those cases where the use of computers for performing assessment is not convenient or possible.

Other authors conducted studies regarding the formative feedback in digital learning environments. In 2013, Narciss (2013) described how the Interactive Tutoring Feedback model could be used in the design and evaluation of strategies of this type. This model describes the interaction between the learner (feedback receiver) and the teacher (feedback source). Espasa et al. (2013) presented a methodological model in order to analyse the interaction of students' groups for improving their essays in online learning environments. Espasa et al.'s model comprises three dimensions (the students' participation, the nature of students' learning, and the quality of students' knowledge) that do not carry the same weight within the model (the students' participation carries less weight). It resulted in improving the online teaching and the learning process. Coll et al. (2013) explored the characteristics of the feedback (focus and type) provided by a teacher and her students inside a collaborative online learning environment. From their results, they found out that the feedback targeted the task and the degree of social participation. They highlighted that this result is in line with the fact that online environments require students to establish ways of interacting among them and to have a keen understanding of the task and its demands rather than focusing on the learning content.

A few comparative studies have also been carried out. Wilson et al. (2011) studied the effectiveness of computer-assisted formative assessment in a large, first-year undergraduate geography course. Statistical analysis showed that the students who used the computer-assisted practice quizzes earned significantly higher grades than those students who did not. Wang (2014) performed a study in which four different e-Learning models were compared (with personalized dynamic assessment, without personalized dynamic assessment, with personalized e-Learning material adaptive annotation, and without personalized e-Learning material adaptive annotation). From their results, the e-Learning models compared without personalized dynamic assessment and the e-Learning models with personalized dynamic assessment were significantly more effective in facilitating student learning achievement and improvement of misconceptions.

In our work, we assume that images can help in the assessment. It is generally accepted that images and graphics can communicate complex ideas with clarity, precision, and efficiency. For example, often the most effective way to describe, explore, and summarize a set of numbers is to look at pictures of those numbers (Tufte, 1989). Reports, executive summaries, and handouts or Power-Point slides used in verbal presentations all benefit from accompanying graphics to capture attention, communicate key information at a glance, and increase understanding and memory retention. Think of graphics as giving the reader the greatest number of ideas, in the shortest time, with the least ink, in the smallest space (Kusek & Rist, 2004; Patton, 1997). It is important to present graphics with written or verbal explanations to ensure their correct interpretation (Torres et al., 2004). Several works have explored the role that images can play in the engagement of schoolchildren. For example, Busschots et al. (2006) explored this aspect for scientific discovery with an astronomy system. They described an online image analysis tool that was developed as part of an interactive, user-centered development of an online system. This system provided a suite of software tools used by schoolchildren and their teachers to study astronomy. In their case,

the astronomical images were spectacular and had the ability to spark the imagination of the participants and, thus, provided a great medium for exploring the role that images can play in the engagement of schoolchildren in scientific discovery. Torres et al. (2004) stated that people learn more when they are engaged with the learning material, when they see, hear, and do something with the content, and when they integrate new knowledge with something they already know. There is also evidence that once an online system has been implemented, over time the response rate will gradually increase (Avery et al., 2006).

Although images are considered important for understanding and solving problems, very few previous works have studied their influence on item solving. One of the works to cite is the study of Dindar et al. (2013). They carried out a study with 112 students in which they compared animated questions vs. static graphic questions. No statistically significant difference was observed in terms of response accuracy between the static group and the animation group. The second work to cite is the study of SAß et al. (2012). They carried out a study with 158 students in which they included or did not include images in the stem and in the answer options. Their results indicated that images in the stem and in the answer options increased the number of correct answers.

III. Material and methods

In our study, as a learning method, we used a computer game that is related to history where the children learned about the different historical ages. In this game, the children travelled through each historical age in this order: Prehistory, Ancient Times, the Middle Ages, the Early Modern Period, and the Contemporary Period. In each historical age, the children learned the main characteristics and events of that historical age. Figure 1 shows graphically the content transmitted by the game. A more detailed explanation about the game used in the study can be found in Martín-SanJosé et al. (2014a, 2014b, 2015).
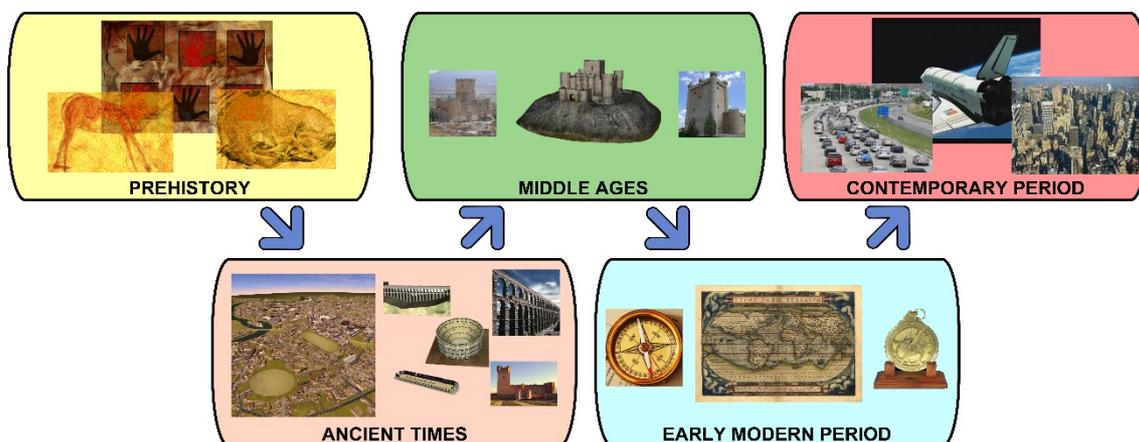

Figure 1. Historical ages that are learned in the game

For the creation of the questionnaires, we used the Gandia Qüest tool, which was developed by the Tesigandia company (http://www.tesigandia.com/en). This tool allows the data of the results to be stored in several formats and also allows data processing. This tool presents a user-friendly interface for creating the forms, and it also makes it easier to add multimedia content to the questionnaires such as images, music, video, and flash applications. The tool can also manage different languages to facilitate the creation of the same form in different languages. In our case, the data was stored in XLSX format (Excel 2010) for the data processing. For two of the questions that required drag and drop interaction, an embedded flash program was also used. In order to maintain data integrity, the data retrieved from

each user was only stored when the whole questionnaire had been completed, otherwise no data was stored.

We used the Gandia Qüest tool, but many other tools can also be used for the same purpose; for example, Website Analysis and MeasureMent Inventory (http://www.wammi.com), Survey Monkey (http://www.surveymonkey.com), Formstack (http://www.formstack.com). Even Google Drive (http://www.drive.google.com) can be used to create a form survey.

IV. Description of the sduty

a. Participants

A total of 94 children participated in the study. There were 46 boys (48.94%) and 48 girls (51.06%). They were between seven and eight years old, and they had already finished their second academic course of primary school. The mean age was 7.56 ± 0.50 years old. The children were students from three different summer schools in Spain.

b. Measurements

To retrieve the data for the analysis, we used three different web-based questionnaires:
1. A text-only questionnaire where all the questions were written in text-only and there were no images on it (Figure 2a).
2. A questionnaire where all the questions had images taken from the game that was played. We refer to the images that appear in the game as real images (Figure 2b). The text was also included.
3. A questionnaire (similar to the previous one), where all the questions had images that did not appear in the game that was played but were representative images of the item specified in the text. We refer to these images as fake images (Figure 2c). The text was also included.

All three questionnaires contained thirteen knowledge questions about the contents of the game, shown in Table 5. A pre-test and a post-test of these three questionnaires were used to carry out the study. We refer to the pre-tests as PreText, PreReal and PreFake; and we refer to the post-tests as PosText, PosReal and PosFake.
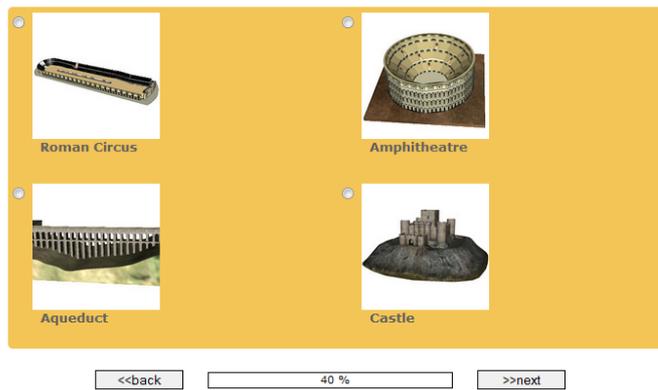

a) Screenshot of Q5 of the text-only questionnaire

5. Where did the gladiators and beasts fight?

Roman Circus

Amphitheatre

Aqueduct

Castle

<<back 40 % >>next

b) Screenshot of Q5 of the questionnaire with real images

5. Where did the gladiators and beasts fight?

Roman Circus

Amphitheatre

Aqueduct

Castle

<<back 40 % >>next
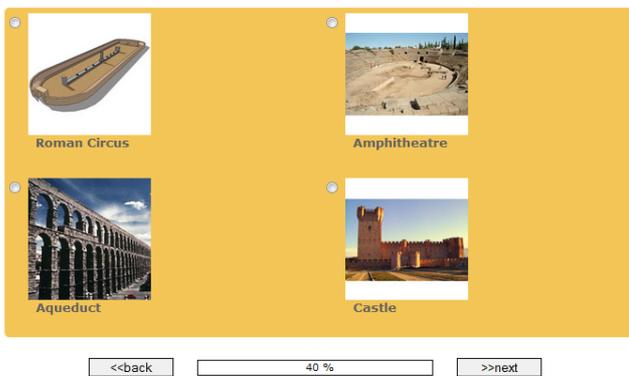
c) Screenshot of Q5 of the questionnaire with fake images

Figure 2. Questionnaire screenshots of Q5

c. Procedure

The participants were assigned to one of the following three groups:
- Group A: The participants who filled out the text-only questionnaires before and after playing the game. There were 36 participants in this group (38.30%).
- Group B: The participants who filled out the questionnaires with real images before and after playing the game. There were 29 participants in this group (30.85%).
- Group C: The participants who filled out the questionnaires with fake images before and after playing the game. There were 29 participants in this group (30.85%).

Figure 3 shows graphically the procedure for the three groups. Since all the questionnaires were filled out online, the answers were automatically stored in a remote database. The questionnaires were filled out individually. Figure 4 shows a child filling out the text-only questionnaire
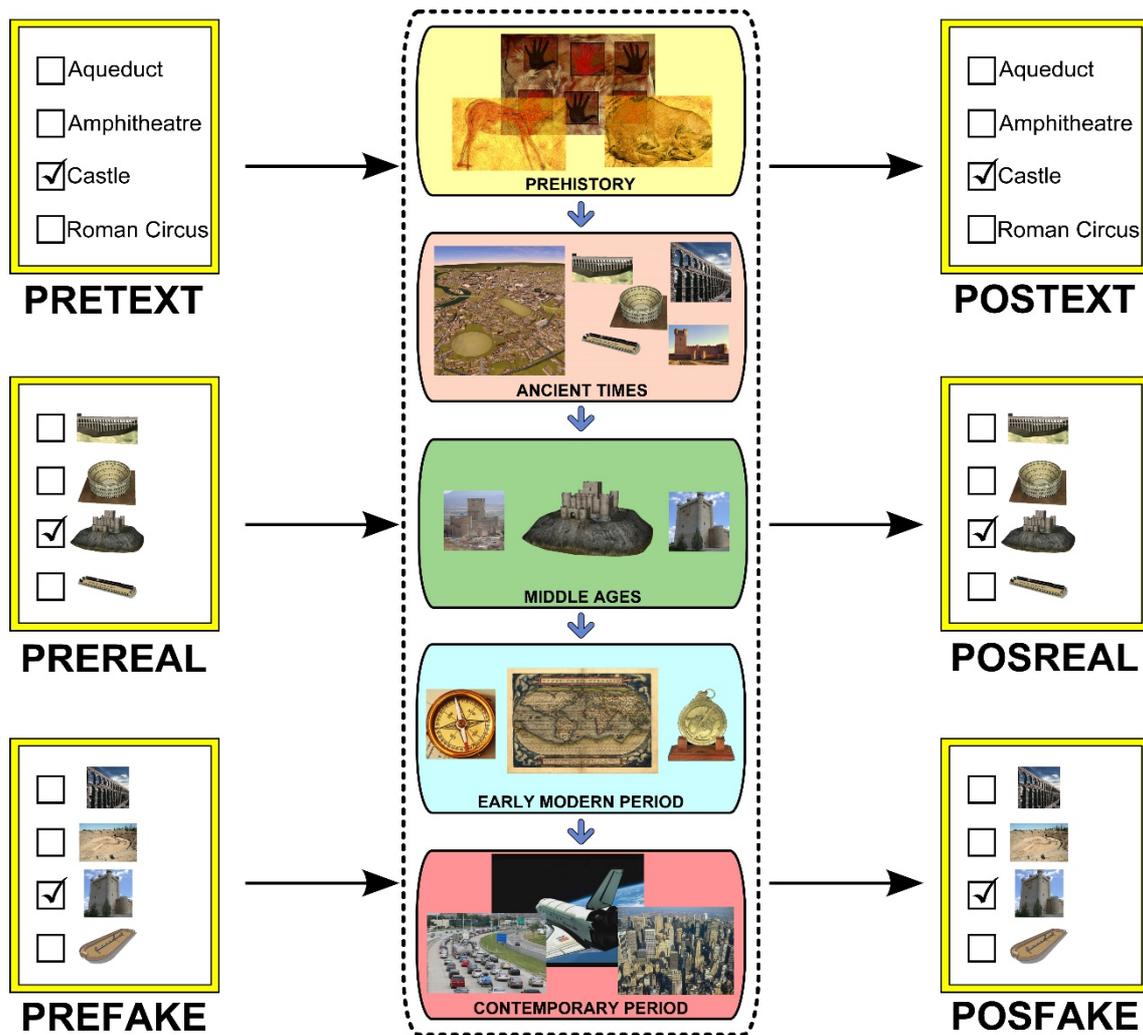
Figure 3. Study procedure

V. Results

a. Learning outcomes

To measure how much the children learned, the knowledge variable was analyzed. This was achieved by analyzing the answers to questions Q1 to Q13 in Table 5 before playing (pre-test) and after playing (post-test). The knowledge value was obtained by summing up all the correct answers. Several t-tests were performed to determine if there were statistically significant differences in the knowledge acquired. Figure 5 shows the box plot for the scores before and after playing the game. As can be observed, there was a high dominance of correct answers after playing the game and using the two questionnaires that had images.
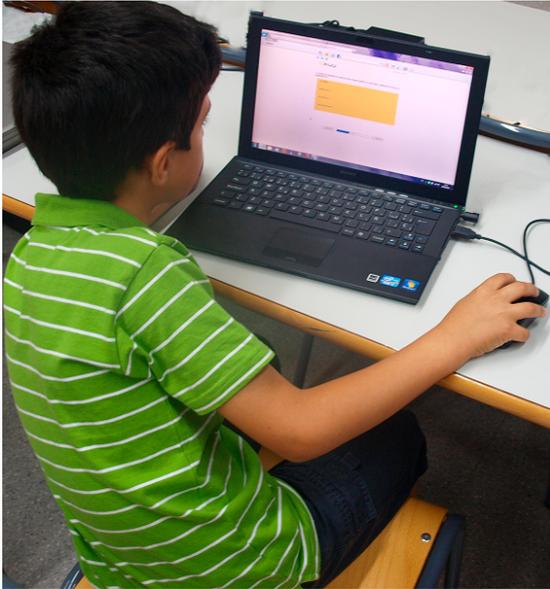
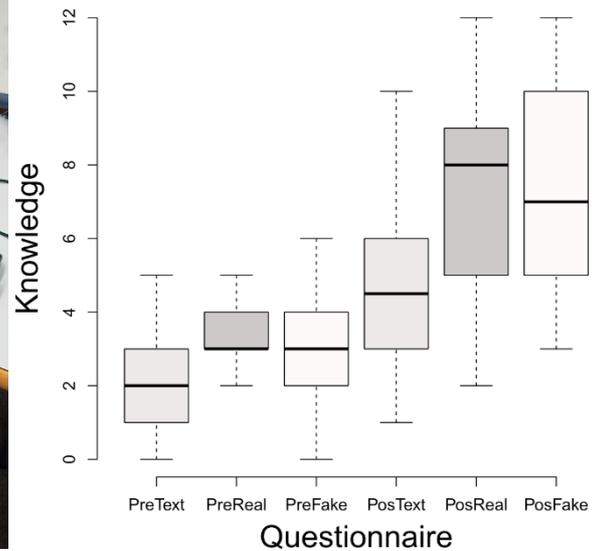Figure 4. A child is filling out the text-only questionnaire

Figure 5. Scores of the knowledge variable before and after playing for each type of questionnaire

All t-tests are shown in the format: (statistic [degrees of freedom], p-value, Cohen's d), and ** indicates the statistical significance at level α=0.05. First, to determine whether or not there were statistically significant differences between the initial knowledge in all types of pre-tests, some un-paired t-tests were performed. Statistically significant differences were found between PreText (2.20 ± 1.50) and PreReal (3.90 ± 1.90) (t[63] = -3.94, p < 0.001**, Cohen's d = 0.98); no statistically significant differences were found between PreFake (3.30 ± 1.60) and PreReal (3.90 ± 1.90) (t[56] = -1.28, p = 0.20, Cohen's d = 0.34). Finally, another unpaired t-test between PreFake (3.30 ± 1.60) and PreText (2.20 ± 1.50) (t[63] = 2.74, p = 0.008**, Cohen's d = 0.68) was performed, where statistically significant differences were found. This proved that children got a better score on the pre-test if it had images (Figure 5). In order to measure the knowledge acquired using each type of questionnaire, several t-tests were performed to compare each pre-test with its post-test. From a paired t-test, the scores of the knowledge variable between PreText (2.20 ± 1.50) and PosText (5.10 ± 2.90) showed statistically significant differences (t[35] = -7.52, p < 0.001**, Cohen's d = 1.25). Another paired t-test between the PreReal (3.90 ± 1.90) and the PosReal (7.40 ± 2.80) questionnaires revealed statistically significant differences (t[28] = -5.85, p < 0.001**, Cohen's d = 1.09). The last comparison between pre-test and post-test was performed between PreFake (3.30 ± 1.60) and PosFake (7.40 ± 2.70) with the results also showing statistically significant differences (t[28] = -8.07, p < 0.001**, Cohen's d = 1.50). These results indicate that regardless of the method used for the assessment, the children acquired knowledge using the game. Finally, in order to determine whether or not there were statistically significant differences between the acquired knowledge in the three groups, further unpaired t-tests were performed between the knowledge in PosText (5.10 ± 2.90) and the knowledge in PosReal (7.40 ± 2.80) (t[63] = -3.36, p = 0.001**, Cohen's d = 0.84) showing that the appearance of the real image helps in choosing the correct answer. When performing this same test using the questionnaire with fake images (7.40 ± 2.70), similar results were obtained (t[63] = 3.35, p = 0.001**, Cohen's d = 0.84) These results showed statistically significant differences. When comparing the two questionnaires that had images, PosFake (7.41 ± 2.65) and PosReal (7.45 ± 2.71), the results showed that there were no statistically significant differences (t[56] = -0.05, p = 0.962, Cohen's d = 0.01). To complete the analysis and check the questions where there were statistically significant differences the following tests were performed. Since the value of the questions were

dichotomous (0, wrong / 1, right), several non-parametric McNemar's tests for paired data were performed for each question between PreText – PosText (Table 1), PreReal

PosReal (Table 2), and PosFake – PosReal (Table 3). Table 1 shows that the children who filled out he text-only questionnaire acquired more knowledge in seven questions. This can be compared with the results in Table 2 provided by the children who filled out the questionnaire with real images of the game. In this case, statistically significant differences were also obtained in seven questions, six of them the same as in the first case. For the children who filled out the questionnaire with fake images, Table 3 shows that there were nine questions with statistically significant differences (including the same six questions as in the previous analyses).

#	PreText	PosText	χ^2	p	φ
Q1	0.05	0.50	14.06	**<0.001****	0.62
Q2	0.08	0.31	6.12	**0.013****	0.41
Q3	0.08	0.61	17.05	**<0.001****	0.69
Q4	0.46	0.33	0.27	0.606	0.09
Q5	0.22	0.33	0.75	0.386	0.14
Q6	0.17	0.22	0.17	0.683	0.07
Q7	0.05	0.22	2.50	0.114	0.26
Q8	0.53	0.64	0.75	0.386	0.14
Q9	0.47	0.56	0.44	0.505	0.11
Q10	0.17	0.44	6.75	**0.009****	0.43
Q11	0.08	0.36	8.10	**0.004****	0.47
Q12	0.05	0.39	8.64	**0.003****	0.49
Q13	0.00	0.17	4.17	**0.041****	0.34

Table 1. Proportions for questions of the PreText and PosText questionnaires, McNemar's test analysis, and φ effect size. *N* = 36

#	PreReal	PosReal	χ^2	p	φ
Q1	0.34	0.72	7.69	**0.006****	0.52
Q2	0.59	0.38	3.12	0.077	0.33
Q3	0.17	0.72	14.06	**<0.001****	0.70
Q4	0.17	0.34	1.23	0.267	0.21
Q5	0.55	0.72	1.45	0.228	0.22
Q6	0.10	0.17	0.17	0.683	0.08
Q7	0.34	0.69	5.79	**0.016****	0.45
Q8	0.62	0.79	1.45	0.228	0.22
Q9	0.48	0.59	0.57	0.450	0.14
Q10	0.21	0.66	7.58	**0.006****	0.51
Q11	0.14	0.66	10.32	**0.001****	0.60
Q12	0.06	0.62	12.50	**<0.001****	0.66
Q13	0.06	0.38	5.82	**0.016****	0.45

Table 2. Proportions for questions of the PreReal and PosReal questionnaires, McNemar's test analysis, and φ effect size. *N* = 28

#	Pre-Fake	Pos-Fake	χ^2	p	φ
Q1	0.28	0.76	10.56	**0.001****	0.60
Q2	0.28	0.59	4.27	**0.039****	0.38

Q3	0.24	0.86	16.06	**<0.001****	0.74
Q4	0.24	0.34	0.36	0.546	0.11
Q5	0.45	0.72	4.90	**0.027****	0.41
Q6	0.06	0.28	3.12	0.077	0.33
Q7	0.14	0.59	9.60	**0.002****	0.58
Q8	0.55	0.76	2.50	0.114	0.29
Q9	0.55	0.72	1.23	0.267	0.21
Q10	0.24	0.62	7.69	**0.006****	0.52
Q11	0.24	0.72	10.56	**0.001****	0.60
Q12	0.00	0.24	5.14	**0.023****	0.42
Q13	0.00	0.21	4.17	**0.041****	0.38

Table 3. Proportions for questions of the PreFake and PosFake questionnaires, McNemar's test analysis, and φ effect size. $N = 28$

In order to compare the acquired knowledge for each question after playing the game, the results between the two post-tests with images were compared with several Fisher exact tests for unpaired data. In this case, only Q12 had statistically significant differences (p = 0.007**) in favor of the questionnaire with real images (proportions 0.62 vs. 0.24).

A multifactorial ANOVA test was also performed to take into consideration several factors simultaneously. The factors were Gender, Age, and Questionnaire. The effect size used was the partial Eta-squared (η^2). The results of the analysis shown in Table 4 indicate that there were statistically significant differences in the Gender and Questionnaire factors. The effect sizes revealed that the most influential factor was the Questionnaire with large size, followed by Gender which had a medium size. No statistically significant differences were found in the interactions between factors. A Tukey's post-hoc pairwise comparison revealed statistically significant differences between PosFake and PosText (p = 0.009**) and between PosReal and PosText (p = 0.007**), which corroborate previous analyses.

Factor	d.f.	F	P	partial η^2
Gender	1	6.99	0.009**	0.078
Age	1	2.87	0.093	0.033
Questionnaire	2	6.91	0.001**	0.144
Interactions	≤ 2	< 2.64	> 0.076	< 0.061

Table 4. Multifactorial ANOVA for the knowledge variable. N = 94

Figure 6 shows the interaction plot between gender and the three types of questionnaires. Boys acquired more knowledge than girls using the Text and Fake-image questionnaires. For the questionnaire with real images, both genders obtained the same score. Figure 7 shows the interaction plot between gender and age, where the older children had higher scores than the younger children. However, this difference between the two ages was not statistically significant.
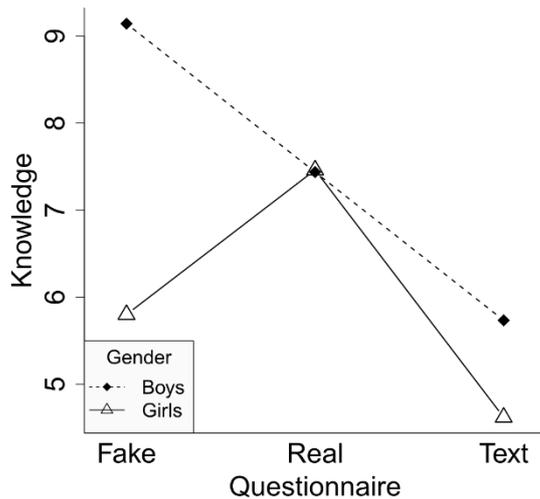
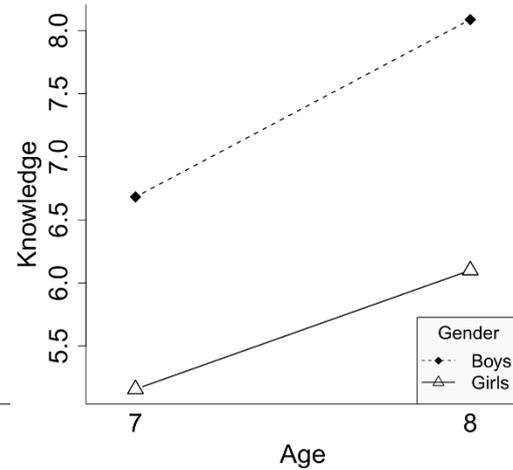Figure 6. Figure 6. Interaction by gender of each questionnaire type

Figure 7. Interaction by gender of each age group

b. Rasch model analysis

To complete the statistical analysis, the dichotomous Rasch model proposed by Georg Rasch was used. This model measures a person's latent trait level from a probabilistic perspective (Rasch, 1960). The probability of a user answering a question correctly depends on the user's underlying ability and the difficulty of question (Fischer, 2006). Figure 8 shows the Item Characteristic Curve (ICC) for every question. The latent dimension shows the ability of the children measured in the interval [-4, 4], with 0 being a child with medium ability. The curve indicates the probability that a child with each ability has to correctly answer a question. The dotted lines represent the medium values of each axis (0 for ability and 0.5 for probability). All the questions in the graph appear ordered by probability to answer the question correctly. Figure 8a shows the ICC for the group of children who used the text questionnaire. It can be observed that, in this group, the hardest question was Q13, where it was necessary for a child to have an ability value of 2 in order to have a probability of 0.5 to answer this question correctly. The easiest question was Q8, where a child with an ability value of -1 was enough to have a probability of 0.5 in order to answer the question correctly. The most balanced question of this group was Q1, which needed an ability of 0 (the medium value) to have a probability of 0.5. Figure 8b shows the ICC for the group of children who used the real-image questionnaire. The order of the questions changed with respect to the previous group. In this group, the most difficult question was Q6 and the easiest was Q8. The most balanced questions for this group were Q1 and Q5 which share the most balanced position. Figure 8c shows the ICC for the group of children who used the fake-image questionnaire. Here the order of the questions also changed. The hardest question was Q13 and the easiest question was Q3. The most balanced questions in this group were Q1 and Q8. In summary, it can be observed from these graphs that, in the text group, the questions are grouped in one cluster. This means that even though the questions have different latent dimensions, they have the same level of magnitude. In contrast, in the other two types of questionnaires, the questions are grouped as two clustered sets of data. This means that the difference between the easier questions and the more difficult questions is more distinguishable in these two types of questionnaires. Although they seem more difficult, the latent dimensions of the children from these groups are enough to solve these questions satisfactorily (Figure 10 completes this information).

a) Text-only
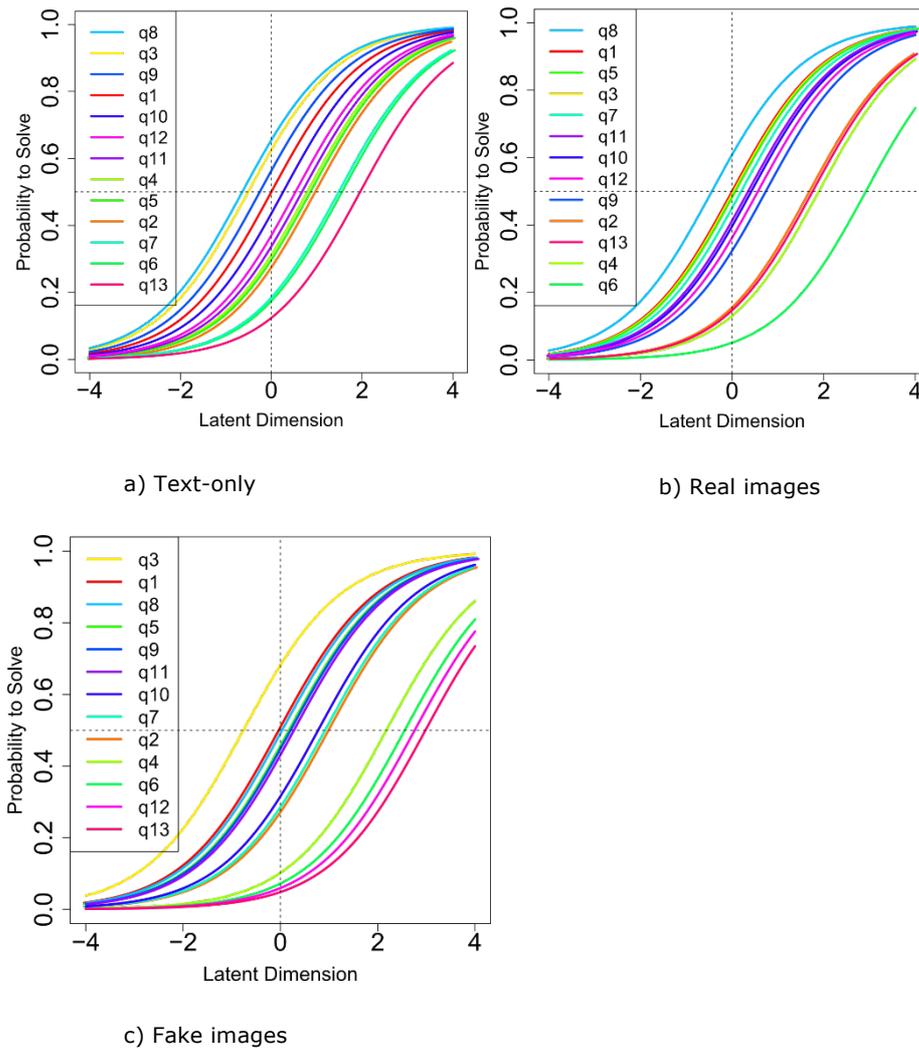
b) Real images

c) Fake images

Figure 8. Item Characteristic Curve (ICC) for all questions

A graphical model check was also performed, where the questions were grouped by raw scores and the ones which were higher than the mean were separated from the ones which were lower. The red lines represent the confidence bands. The results for the questions are shown in the graphs in Figure 9. For the group of children who used the text questionnaires (Figure 9a), it can be observed that only Q2 is narrowly out of the confidence bands; for the group of children who used the real-image questionnaire (Figure 9b), Q10 is touching the confidence bands; finally, for the group of children who used the fake-image questionnaire (Figure 9c), every question is inside the confidence bands. Therefore, the questions are appropriate for the assessment of the acquired knowledge for the three types of questionnaires.

Graphical Model Check

a) Text-only

Graphical Model Check
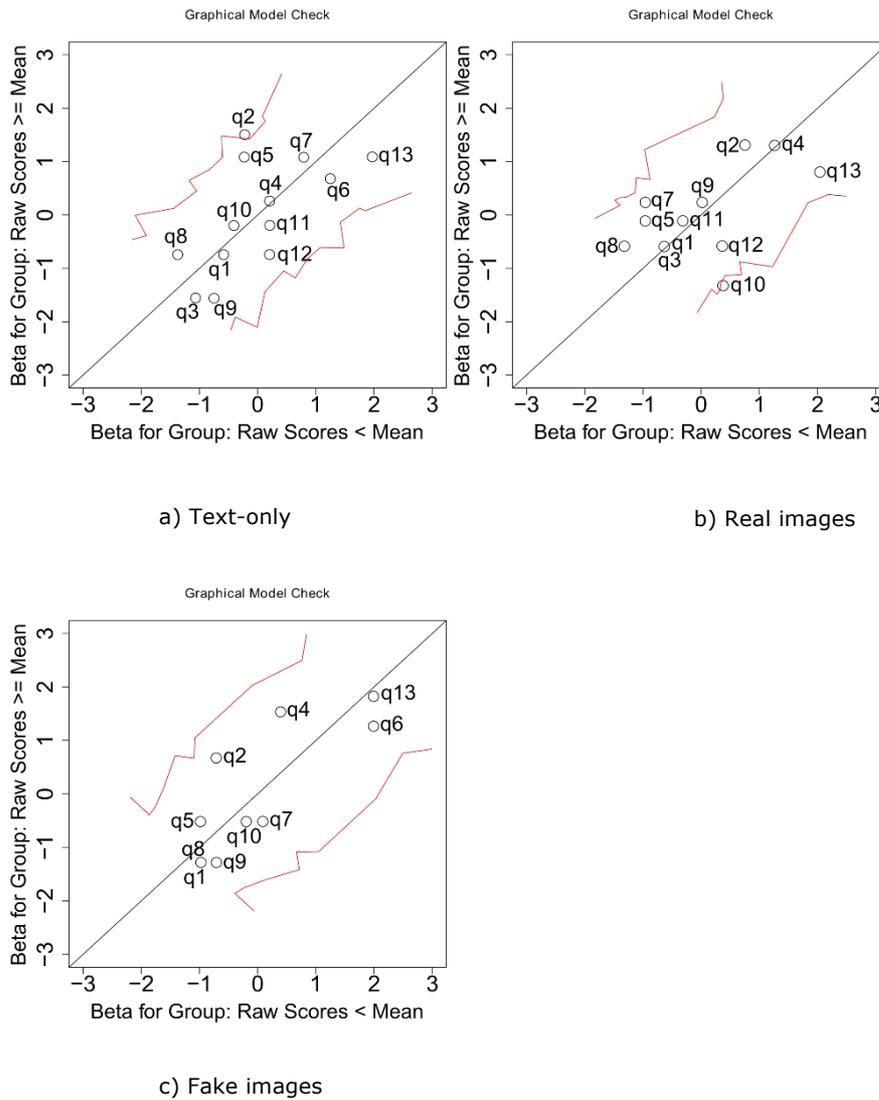
b) Real images

Graphical Model Check

c) Fake images

Figure 9. Graphical model check

In order to visually check the children and the questions, a Person-Item Map was plotted, where the estimated ability of the child and the question difficulty measures are placed side by side in one vertical dimension. The questions appear in order of difficulty. The Person-Parameter Distribution (which is at the top of the graph) is a distribution of the children's abilities.

The Person-Item Map for each group of children is shown in Figure 10. For the text questionnaire group (Figure 10a), the hardest question (Q13) was easier than the ability of 8.33% of the children, and the easiest question (Q8) was more difficult than the ability of 33.33% of the children. For the real-image questionnaire group (Figure 10b), the hardest question (Q6) was easier than the ability of 6.89% of the children, and the easiest question (Q8) was more difficult than the ability of 10.34% of the children. For the fake-image questionnaire group (Figure 10c), the most difficult question (Q13) was easier than the ability of 10.34% of the children, and the easiest question (Q3) was harder than the ability of 0% of the children.

It can be observed in Figure 10 that the distributions of children who used a questionnaire with images are moved to the right, which means that most of the children were able to correctly answer most of the questions. In the case of the text-only questionnaire, the distribution shows that most of the children were in the lower levels and near the easiest questions. The questions were grouped in the same way they were distributed in Figure 8. In the text-only group, the questions were grouped in one cluster, and in the other two questionnaires, there were two differentiate clusters of questions. In summary, it can be concluded from these graphs that the children who used the questionnaires with images acquired greater latent ability for answering the questions than the children who used the text-only questionnaire.

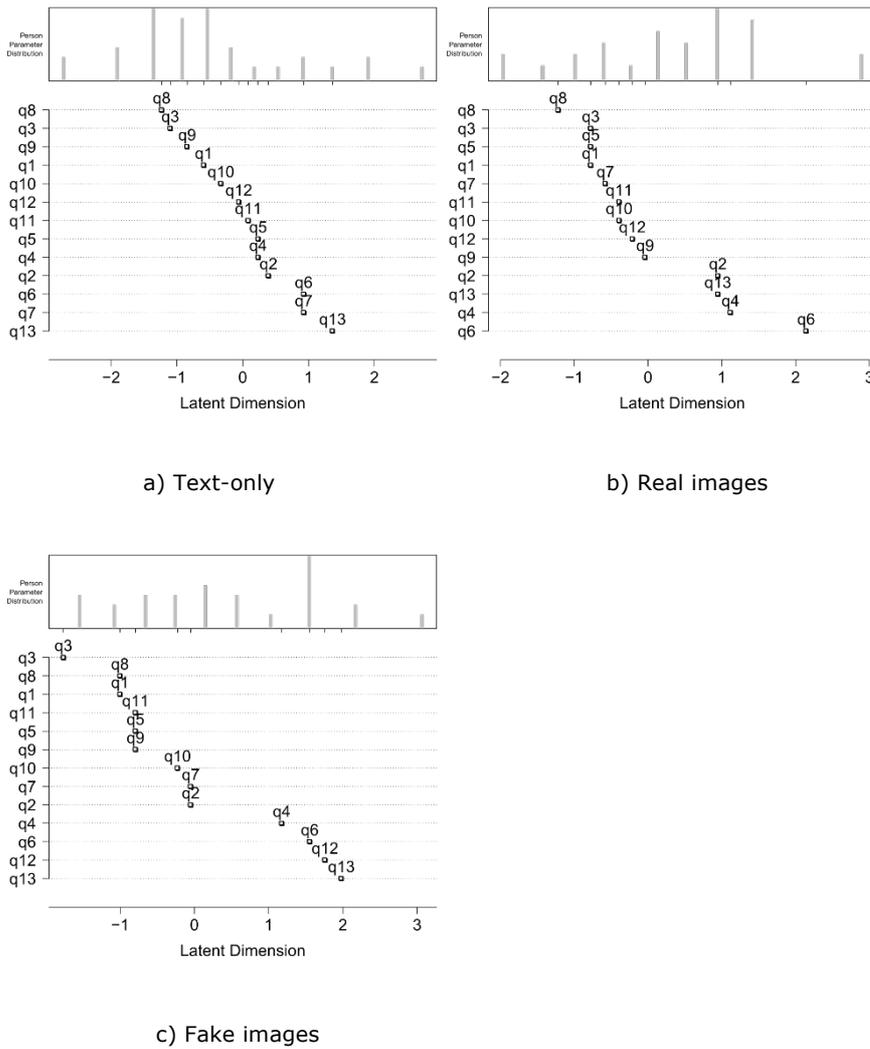a) Text-only

b) Real images

c) Fake images

Figure 10. Person-Item Map

To check the goodness of fit of the Rasch model, the test proposed by Andersen (1973) was used. This test is based on a comparison between the difficulties estimated from different score groups and estimates, resulting in a conditional likelihood ratio. Andersen stated that 2 times the logarithm of this ratio is χ^2-distributed when the Rasch model is true. In our study, this test offered the following values that fit the Chi-squared distribution: LRvalue = 14.44, df = 12, p = 0.274. Therefore, the Rasch model is true in our study.

VI. Conclusions

Three different types of questionnaires were designed and tested for assessment purposes. We carried out a study with children to determine whether the use of images that accompany an item to be identified in a question affects the selection of the correct answer in any way. We compared the text-only questionnaires with questionnaires that had images that appeared in the game that was played (real images) and questionnaires that had representative images of the item specified (fake images) that had not appeared in the game played.

From the initial knowledge of the three groups, statistically significant differences were found when comparing the text-only questionnaire with either of the two questionnaires with images. No statistically significant differences were found between the questionnaires with real and fake images. This was the result that we expected since, before playing the game, the two types of images (real or fake) represent the same concept. This result implies that the images gave an additional clue in selecting the correct answer

Even though it was not the primary objective of the study, the acquired knowledge variable was analysed to assure that the learning method used is effective when it comes to transmitting knowledge in the short-term. The results indicated that regardless of the questionnaire used for the assessment (text-only, real, or fake), the children acquired statistically significant improvement in knowledge using the game. Therefore, the game used is an effective learning method.

From the knowledge scores obtained after playing the game, statistically significant differences were found only when comparing the text-only questionnaire with either of the two questionnaires with images regardless of whether or not the images were exactly the same as the ones used in the game. These results corroborate the primary hypothesis (the questionnaire with images are better than the text-only questionnaire) but do not support the secondary hypothesis (real images are better than fake images). Even though we expected both hypotheses to be corroborated, it is still an excellent result because it means that images (real or fake) help the students to choose the right answer. As in the pre-test, the students did not choose the right answer with only text. However, when the associated image was included, they were able to choose the right answer. Therefore, it can be concluded that, to a great extent, the use of images in the questionnaires helps student to select the correct answer. This conclusion is in line with the work of SAß et al. (2012). Moreover, our work also demonstrates that it does not matter whether or not images are used during the instruction of the material or whether the images used were the same or different as those used during instruction. In formative assessment, the training does not finish until the end of the course and the assessment is part of the training process. If images are added to answers, the children can relate an image with its definition during the assessment, which contributes to completing their training. Therefore, based on our results, images added to answers could be used in formative assessment as a reinforcement of the knowledge that the children have while performing the tests.

Based on our own studies and those of other authors mentioned in this work, we can conclude that computer-based assessment offers different advantages. CBA helps in the increasing of the engagement of the students (Anderson et al., 2005). CBA reduces the costs of paper, time, and processing, and is less vulnerable to the influence of the faculty (Dommeyer et al., 2002b), and CBA facilitates self-assessment. Self-assessment has advantages for both teachers and students (McConnell, 2006). It provides immediate feedback and helps to eliminate the distance between teachers and students. Moreover, students are more independent, which can promote self-confidence. If the assessment is online, it helps in overcoming the problems that traditional learning environments have

(restrictions of teaching schedules and large numbers of concurrent students) (Wang, 2011; Wang et al., 2007).

Our study provides many possible options for further research. One way of guiding students in their learning that is normally used by teachers is to use instructional prompts when students give an incorrect answer. If the feedback arrives via a graduated prompt approach, it facilitates the students' thinking and gives correct answers step by step (Campione & Brown, 1987). Our proposal could be incorporated in systems that already include feedback in order to determine to what extent the inclusion of images improves self-assessment. The same idea could be applied to systems that include personalized assessment. According to Wang (2014), learners are likely to experience better e-Learning effectiveness when they conduct self-evaluation via Web-based dynamic assessment.

With regard to the factors that influence the Behavioural Intention to Use a computer-based assessment, Terzis and Economides (2011) conducted a study to investigate these factors. From their results, they concluded that Perceived Ease of Use and Perceived Playfulness have a direct effect on the use of computer-based assessment. According to those in charge of our study who were supervising the activities, the children had no problems using the questionnaires. Informal questions and the children's comments indicate that it was easy to use. However, a formal study could confirm this assertion and also take into account the playfulness aspect.

In this work, we have compared three types of questionnaires, but other comparisons are also possible for future studies; for example, using only images without text; using mobile devices vs. PCs for filling out the questionnaires; checking whether the use of images offers similar results for adults, and also different academic subjects. Finally, we hope that our study contributes to the effectiveness of formative assessment in general and formative assessment specifically for children

Acknowledgements

References

Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, *38*(1), 123–140.

Anderson, H. M., Cain, J., & Bird, E. (2005). Online Student Course Evaluations: Review of Literature and a Pilot Study. *American Journal of Pharmaceutical Education*, *69*(1), 34–43, article no. 5. doi:10.5688/aj690105

Arreola, R. A. (1995). *Developing a Comprehensive Faculty Evaluation System*. Boston: Jossey-Bass.

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic Course Evaluations: Does an Online Delivery System Influence Student Evaluations? *The Journal of Economic Education*, *37*(1), 21–37. doi:10.3200/JECE.37.1.21-37

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: brain, mind, experience, and school* (expanded.). Washington, DC: National Academy Press.

Brown, J. E., Hinze, S. R., & Pellegrino, J. W. (2008). Technology and formative assessment. In T. L. Good (Ed.), *Technology* (pp. 245–255). Thousand Oaks, California: Sage Publications.

Busschots, B., Raeside, L., Keating, J. G., & Waddington, S. (2006). The VTIE Collaborative Writing Environment. In *WEBIST conference on web information systems and technologies* (pp. 221–228).

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: An international approach to evaluating learning potential* (pp. 82–115). New York: The Guilford Press.

Challis, D. (2005). Committing to quality learning through adaptive online assessment. *Assessment & Evaluation in Higher Education*, *30*(5), 519–527. doi:10.1080/02602930500187030

Cisco, Intel, & Microsoft. (2009). Transforming education: Assessing and teaching 21st century skills. Retrieved from http://download.microsoft.com/download/

Coll, C., Rochera, M. J., Gispert, I. de, & Barriga, F. D. (2013). Distribution of Feedback among Teacher and Students in Online Collaborative Learning in Small Groups. Digital Education Review, (23), 27–45.

Conejo, R., Barros, B., Guzmán, E., & Garcia-Viñas, J.-I. (2013). A web based collaborative testing environment. *Computers & Education*, *68*, 440–457. doi:10.1016/j.compedu.2013.06.001

Dindar, M., Yurdakul, I. K., & Dönmez, F. I. (2013). Multimedia in Test Items: Animated Questions vs. Static Graphics Questions. *Procedia - Social and Behavioral Sciences*, *106*, 1876–1882. doi:10.1016/j.sbspro.2013.12.213

Dommeyer, C. J., Baum, P., Chapman, K. S., & Hanna, R. W. (2002a). Attitudes of Business Faculty Towards Two Methods of Collecting Teaching Evaluations: Paper vs. Online. *Assessment & Evaluation in Higher Education*, *27*(5), 455–462. doi:10.1080/0260293022000009320

Dommeyer, C. J., Baum, P., & Hanna, R. W. (2002b). College Students' Attitudes Toward Methods of Collecting Teaching Evaluations: In-Class Versus On-Line. *Journal of Education for Business*, *78*(1), 11–15. doi:10.1080/08832320209599691

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*(5), 611–623. doi:10.1080/02602930410001689171

Dube, T., & Ma, M. (2010). A flexible e-assessment system inspired by design methodology management to accommodate diverse learning styles. *International Journal of Digital Society*, *1*(1), 44–52.

Espasa, A., Guasch, T., & Alvarez, I. M. (2013). Analysis of Feedback Processes in Online Group Interaction: a Methodological Model. Digital Education Review, (23), 59–73.

Fischer, G. H. (2006). Rasch Models. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26, pp. 515–585). Handbook of Statistics. Elsevier. doi:10.1016/S0169-7161(06)26016-4

Gathy, P., Denef, J.-F., & Haumont, S. (1991). Computer-assisted self-assessment (CASA) in histology. *Computers & Education*, *17*(2), 109–116. doi:10.1016/0360-1315(91)90089-A

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351. doi:10.1016/j.compedu.2011.06.004

Hargreaves, E. (2008). Assessment. In G. McCulloch & D. Crook (Eds.), *The Routledge international encyclopedia of education* (pp. 37–38). New York: Routledge.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. doi:10.3102/003465430298487

Heinrich, E., & Lawn, A. (2004). Onscreen marking support for formative assessment. In *Proceedings from Ed-Media 2004 world conference on educational multimedia, hypermedia & telecommunications* (pp. 1985–1992).

Hwang, G.-J., & Chang, H.-F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*, *56*(4), 1023–1031. doi:10.1016/j.compedu.2010.12.002

IMS Global Learning Consortium. (2008). IMS question and test interoperability specification. Retrieved from http://www.imsglobal.org/question

Keeves, J. P. (1994). Assessment in schools, methods of assessment. In Husen, Torsten, Postlethwaite, & Neville (Eds.), *The international encyclopedia of education* (2nd ed., pp. 362–370). Oxford: Pergamon Press.

Kusek, J. Z., & Rist, R. C. (2004). *Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners*. Washington, DC: World Bank Publications.

Layne, B. H., Decristoforo, J. R., & Mcginty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, *40*(2), 221–232. doi:10.1023/A:1018738731032

Lazarinis, F., Green, S., & Pearson, E. (2010). Creating personalized assessments based on learner knowledge and objectives in a hypermedia Web testing application. *Computers & Education*, *55*(4), 1732–1743. doi:10.1016/j.compedu.2010.07.019

Llamas-Nistal, M., Caeiro, M., Castro, M., Plaza, I., & Tovar, E. (2011). Use of LMS functionalities in engineering education. In *Proceedings from 41st ASEE/IEEE frontiers in education conference (FIE 2011)* (pp. S1G–1–S1G–6, 12–15).

Llamas-Nistal, M., Fernández-Iglesias, M. J., González-Tato, J., & Mikic-Fonte, F. A. (2013). Blended e-assessment: Migrating classical exams to the digital world. *Computers & Education*, *62*, 72–87. doi:10.1016/j.compedu.2012.10.021

Martín-SanJosé, J.-F., Juan, M.-C., Gil-Gómez, J.-A., & Rando, N. (2014a). Flexible learning itinerary vs. linear learning itinerary. *Science of Computer Programming*, *88*, 3–21. doi:10.1016/j.scico.2013.12.009

Martín-SanJosé, J.-F., Juan, M.-C., Seguí, I., & García-García, I. (2015). The effects of computer-based games and collaboration in large groups vs. collaboration in pairs or traditional methods. *Computers & Education*, *87*, 42–54. doi:10.1016/j.compedu.2015.03.018

Martín-SanJosé, J.-F., Juan, M.-C., Torres, E., & Vicent, M. J. (2014b). Playful interaction for learning collaboratively and individually. *Journal of Ambient Intelligence and Smart Environments*, *6*(3), 295–311. doi:10.3233/AIS-140257

McConnell, D. (2006). *E-Learning groups and communities: Imaging learning in the age of the internet*. OU Press.

Monahan, T., McArdle, G., & Bertolotto, M. (2008). Virtual reality for collaborative e-learning. *Computers & Education*, *50*(4), 1339–1353. doi:10.1016/j.compedu.2006.12.008

Narciss, S. (2013). Designing and Evaluating Tutoring Feedback Strategies for digital learning environments on the basis of the Interactive Tutoring Feedback Model. Digital Education Review, (23), 7–26.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. doi:10.1080/03075070600572090

Oosterhof, A., Conrad, R. M., & Ely, D. P. (2008). *Assessing learners online*. New Yersey: Pearson.

Patton, M. Q. (1997). *Utilization-Focused Evaluation* (3rd ed.). Thousand Oaks, California: Sage Publications.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rashad, A. M., Youssif, A. A., Abdel-Ghafar, R. A., & Labib, A. E. (2008). E-assessment tool: a course assessment tool integrated into knowledge assessment. In M. Iskander (Ed.), *Innovative techniques in instruction technology, e-learning, e-assessment, and education* (pp. 7–12). Netherlands: Springer.

Ravelli, B. (2000). Anonymous Online Teaching Assessments: Preliminary Findings. In *Annual National Conference of the American Association for Higher Education*. Charlotte, North Carolina.

Reeves, T. C., & Hedberg, J. G. (2009). Evaluation strategies for open and distributed learning environments. In C. Spratt & P. Lajbcygier (Eds.), *E-Learning technologies and evidence based assessment approaches* (pp. 234–253). New York: Information Science Reference.

Rota, D. R. (1981). *Computer assisted instruction, lecture instruction, and combined computer-assisted/lecture instruction: A comparative experiment*. Unpublished doctoral dissertation . University of Pittsburgh.

Saß, S., Wittwer, J., Senkbeil, M., & Köller, O. (2012). Pictures in Test Items: Effects on Response Time and Response Correctness. *Applied Cognitive Psychology*, *26*(1), 70–81. doi:10.1002/acp.1798

Sorenson, D. L., & Johnson, T. D. (2003). *Online Student Ratings of Instruction: New Directions for Teaching and Learning*. San Francisco, CA: Jossey-Bass.

Stankov, S., Grubišić, A., & Žitko, B. (2004). E-learning paradigm and Intelligent tutoring systems. In Z. Kniewald (Ed.), *Annual 2004 of the Croatian academy of engineering* (pp. 21–31). Zagreb: Croatian Academy of Engineering.

Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer based assessment. *Computers & Education*, *56*(4), 1032–1044. doi:10.1016/j.compedu.2010.11.017

Torres, R. T., Preskill, H. S., & Piontek, M. E. (2004). *Evaluation Strategies for Communicating and Reporting: Enhancing Learning in Organizations* (2nd ed.). Thousand Oaks, California: Sage Publications.

Tufte, E. R. (1989). *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.

Vonderwell, S., Liang, X., & Alderman, K. (2007). Asynchronous discussions and assessment in online learning. *Journal of Research on Technology in Education*, *39*(3), 309–328.

Wang, T.-H. (2010). Web-based dynamic assessment: Taking assessment as teaching and learning strategy for improving students' e-Learning effectiveness. *Computers & Education*, *54*(4), 1157–1166. doi:10.1016/j.compedu.2009.11.001

Wang, T.-H. (2011). Developing Web-based assessment strategies for facilitating junior high school students to perform self-regulated learning in an e-Learning environment. *Computers & Education*, *57*(2), 1801–1812. doi:10.1016/j.compedu.2011.01.003

Wang, T.-H. (2014). Developing an assessment-centered e-Learning system for improving student learning effectiveness. *Computers & Education*, (published online). doi:10.1016/j.compedu.2013.12.002

Wang, Y.-S., Wang, H.-Y., & Shee, D. Y. (2007). Measuring e-learning systems success in an organizational context: Scale development and validation. *Computers in Human Behavior*, *23*(4), 1792–1808. doi:10.1016/j.chb.2005.10.006

Wilson, K., Boyd, C., Chen, L., & Jamal, S. (2011). Improving student performance in a first-year geography course: Examining the importance of computer-assisted formative assessment. *Computers & Education*, *57*(2), 1493–1500. doi:10.1016/j.compedu.2011.02.011

Appendix

This appendix presents all the knowledge questions that were used in this study. The choices to be selected as answers are placed below the questions. The column labeled with # shows the question numbering.

#	Question
Q1	Which of the following figures did the cavemen paint in the caves? a) Houses b) Deers c) Bisons d) Boats e) Hands . f) Carts
Q2	Tell the name of a cave with cave paintings a) Bajamira cave b) Miradentro cave c) Altamira cave d) Cave paintings cave
Q3	Which of the following colours were used for painting in Prehistory? a) Green b) Red c) Violet d) Blue e) Ochre f) Black
Q4	Ancient Times started with the: a) Invention of the wheel b) Invention of writing c) Discovery of America d) Fall of the Roman Empire e) Invention of the compass
Q5	Where did the gladiators and beasts fight? a) Roman circus b) Aqueduct c) Amphitheatre d) Castle
Q6	Which of the following characteristics correspond to Ancient Times? a) Some people lived in castles

b)   There were aqueducts and amphitheatres

c)   Mankind started to paint in caves

d)   The compass was used to navigate.

Q7   What is the name of the fortification in front of the walls of the castle that protected the main door from enemies?

a) Moat   b) Keep

c) Barbican   d) Defensive tower

Q8   Which structure surrounds the castle and can be full of water?

a) Barbican   b) Moat

c) Road   d) Keep

Q9   What part of the castle did the Castle's Lord and his family live in?

a) Keep   b) Barbican

c) Wall   d) Defensive tower

Q10   Which event marked the start of the Early Modern Period?

a)   The invention of writing

b)   The discovery of America

c)   The invention of the mobile phone

d)   The trip to the moon

Q11   Select the inventions used for sailing in the Early Modern Period

a) Compass   b) Television   c) Astrolabe

d) Map   e) Mobile phone   f) Spaceship

Q12   Place the historical ages in the correct order

a) Ancient Times   b) Contemporary Period

c) Prehistory   d) The Early Modern Period

e) The Middle Ages

Q13   Place each invention in the correct historical age

a) Map   b) Mobile phone   c) Cave paintings

d) Aqueduct   e) Castle

Table 5. Learning questions numbered as in the questionnaires

Recommended citation

Martín-SanJosé, J.F., Juan, M.C., Vivó, R., Abad, F. (2015) The Effects of Images on Multiple-choice Questions in Computer-based Formative Assessment. In: *Digital Education Review, 28*, 123-144. [Accessed: dd/mm/yyyy] http://greav.ub.edu/der

Subscribe & Contact DER

In order to subscribe to DER, please fill the form at http://greav.ub.edu/der