Automatic Test Item Creation in Self-Regulated Learning
Evaluating Quality of Questions in a Latin American Experience

Gudrun Wesiak[1], Rocael Hernández Rizzardini[2], Hector Amado-Salvatierra[2], Christian Guetl[1,3] and Mohammed Smadi[1]

[1]*IICM, University of Technology, Graz, Austria*
[2]*GES department, Galileo University, Guatemala, Guatemala*
[3]*Curtin University, Perth, Western Australia, Australia*

Keywords: Self-Regulated Learning, Automatic Test Item Generation, Evaluation Study, e-Assessment.

Abstract: The research area of self-regulated learning (SRL) has shown the importance of the learner's role in their cognitive and meta-cognitive strategies to self-regulate their learning. One fundamental step is to self-assess the knowledge acquired, to identify key concepts, and review the understanding about them. In this paper, we present an experimental setting in Guatemala, with students from several countries. The study provides evaluation results from the use of an enhanced automatic question creation tool (EAQC) for a self-regulated learning online environment. In addition to assessment quality, motivational and emotional aspects, usability, and tasks value are addressed. The EAQC extracts concepts from a given text and automatically creates different types of questions based on either the self-generated concepts or on concepts supplied by the user. The findings show comparable quality of automatically and human generated concepts, while questions created by a teacher were in part evaluated higher than computer-generated questions. Whereas difficulty and terminology of questions were evaluated equally, teacher questions where considered to be more relevant and more meaningful. Therefore, future improvements should especially focus on these aspects of questions quality.

1 INTRODUCTION

Learners are increasingly faced with a huge amount of knowledge and with new learning tasks that require abilities to improve the organization of their learning process. Therefore there is a shift from learning controlled by the teacher to a process where the students regulate themselves (Kroop et al., 2012). There is a relevant amount of research effort on student's self-regulated learning (SRL) strategies, mostly focused on highly controlled learning environments such as intelligent systems for tutoring, self-reflection, formative assessment and feedback (Zimmerman, 1989; Bannert, 2006; Nicol and Macfarlane-Dick, 2006; Conati and Vanlehn, 2000). Also there is a need to foster the students' SRL skills when they are not able to predict the output of a learning activity or the best learning path in virtual learning environments. Good SRL skills will help to adapt the learning process and improve the learning outcome. Therefore providing tools for continuous assessment and feedback is a key for the learning process. In this sense, in their introduction to assessment as a tool for learning Dochy and McDowell (1997) already pointed out how important assessment and evaluation are at all stages of a learning process. For a deeper understanding of a learning content reflection, feedback, learning path and an integration of learning and assessment are crucial elements. In SRL students often perform individual online searches regarding some learning topic and therefore their learning resources differ from those of their peers. In general, learning materials found in the web do not have integrated assessment tools (such as quizzes or short knowledge tests) and students are therefore not able to get feedback on their understanding of the materials. With an automatic question creation tool for natural texts, learners receive a possibility to generate their own little quizzes for any textual learning resource found in the web - independent of time, place, or the input of a teacher or tutor. Thus, they can deepen their understanding and learn more effectively by answering questions and obtaining

Wesiak G., Hernandez Rizzardini R., Amado-Salvatierra H., Guetl C. and Smadi M..
Automatic Test Item Creation in Self-Regulated Learning - Evaluating Quality of Questions in a Latin American Experience.
DOI: 10.5220/0004387803510360
In *Proceedings of the 5th International Conference on Computer Supported Education* (CSEDU-2013), pages 351-360
ISBN: 978-989-8565-53-2

feedback to their chosen learning context.

There has been extensive work on the idea of automatic creation of assessment items. One key component is the extraction and identification of the most relevant concepts used in natural language texts of the learning contents, still being a current research focus (Villalon and Calvo, 2009). There are a variety of experiences for automatically or semi-automatically generated test items from a given input text (e.g. Stanescu et al., 2008; Liu and Calvo, 2009). As an example, Liu and Calvo (2009) provided a tool that is capable to generate open-ended questions out of essays. Chen, Aist and Mostow (2009) created a tool that generates also open-ended questions but from informational texts. According to Agarwal, Shah, and Mannem (2011) the two main challenges in automatic test item generation are (a) to identify a content for which an item should be created and (b) to find a corresponding test-item type. Most research in this field deals with only one type of test item (mostly open-ended or multiple choice) generated for one specified content, e.g. a given sentence (Goto et al. 2010). Exceptions are for example Brown, Frishkoff, and Eskenazi (2005), who generated multiple choice and assignment items for vocabulary assessment or Myller (2007), who used multiple choice, single-choice, and open ended questions in his work on prediction questions for visualizations. Guetl, Lankmayr, Weinhofer, & Hoefler (2011) have developed a tool which generates four types of questions out of natural text.

In our proposed approach, a learning environment scenario enables students to read a text, then select key concepts from this text, and finally get an automatic assessment that will help to improve their understanding and knowledge acquisition. The test items are either based on the concepts selected by the student or on concepts extracted by the tool itself. We used the EAQC by Guetl et al., (2011) integrated in an online learning environment (Intelligent Web Teacher, IWT, by Capuno et al., 2009) to generate test items for two different topics. EAQC has already been evaluated with several quality criteria (Hoefler et al., 2011; 2012). However, previous studies were set up as stand-alone experiences with undergraduate students as participants and no direct involvement of the teacher. To test the quality of the EAQC in a more realistic, broader, and more discerning setting, this study was carried out in Guatemala with participants from different countries in Latin America. Students were enrolled in a full online SRL course and are for the most part teachers with different cultural background. Furthermore, EAQC test items were compared to items generated by participants' actual teacher. The main goal of the study was to evaluate the quality of the concepts and questions generated with the EAQC.

This paper is organized as follows: after a description of the EAQC and IWT tools, the research methodology including participants, instruments used, and the experimental design, is presented. The next section reports the results, which are followed by a discussion and final conclusions.

2 THE ENHANCED AUTOMATIC QUESTION CREATION TOOL (EAQC)

The Enhanced Automatic Question Creator EAQC (Guetl et al., 2011) is a tool that automatically creates test items from textual learning material. EAQC has the functionality to generate different types of test items from textual input. The item types supported by the EAQC are: open end (OE), multiple choice (MC), true or false (TF), and fill-in-the-blank (FiB) questions. EAQC processes the textual input, which can come in a diversity of file formats, then extracts the most important content from the text provided and performs a relation of concepts. EAQC creates the different test items and also creates reference answers. The EAQC architecture provides a flexible extension to multiple languages, an important feature to test the tool with international scenarios. Furthermore, the EAQC is capable to export the test items in a standard compliant format (e.g. IMS Question and Test Interoperability QTI).

EAQC supports mainly the following scenarios: First, a totally automatic question creation scenario where students and teachers cannot control the assessment authoring but they only select the learning materials. Second a more interactive setting where students and teachers can select not only the learning material but also they can tag and select concepts using an online editor. These concepts are used for creating the questions, which finally results in an assessment that has been created automatically, but is based on users' selection of relevant concepts.
The generation of questions in EAQC can be divided into two processes. In a first step, the EAQC extracts concepts out of the text, which can be viewed by the user. In a second step, questions are generated based on the extracted concepts. Therefore, the user can choose which concepts are to be used for the

generation of questions, and which types of questions are to be generated (e.g. two multiple choice, three true/false questions, etc.). The generated questions are presented as test to the user and right after taking the test, students receive feedback on their performance. Thereby, the EAQC lists again all questions but this time with the answer given by the student, the correct answer, and the received points.

2.1 The e-Learning Platform (IWT)

The EAQC was used by the students within the Intelligent Web Teacher (IWT) platform. IWT provides flexibility and extensibility characteristics for contents and services at a low level and for strategies and models at a higher level (Capuano et al., 2009). Furthermore, IWT platform provides easy to adopt didactic experiences based on user preferences for a personalized learning.

The EAQC was integrated into the IWT after previous studies (Al-Smadi et al., 2011; Hoefler et al., 2011; 2012) and further improved with two new features. These features were the implementation of a function for adding concepts and other function for tagging concepts. To add new concepts in the tool, the user open the list of concepts the EAQC extracted from a text and then is chooses further concepts from a list of words and phrases contained in the text. Furthermore, the user can order the final list by relevance and choose only the most relevant concepts to generate questions. The function for tagging concepts allows students to simply highlight a concept within the text and then save it to their concept list. Afterwards, they can generate questions on the basis of their self-extracted concepts.

3 RESEARCH METHODOLOGY

3.1 Motivation and Goals

The main goals of this study were (a) to systematically test the quality of concepts and questions generated with EAQC by comparing them to those generated by a real teacher (with course material used in an actual learning setting), and (b) to provide an automatic assessment tool that motivates and supports students in a SRL environment.

3.2 Participants

The experiment was carried out in the Institute Von Neumann (IVN) of Galileo University, Guatemala. IVN is an online higher education institute.

Thirty students enrolled in a course on "Learning models and processes for e-Learning" participated in the study. From these students, 27 were from Guatemala, 2 from the United States, and 1 from Colombia. The course is part of a complete online learning Master degree program in "e-Learning Management and Production". The students are university professors, consultants and instructors at corporations.

Twelve students were male and 18 were female. Participants were between 22 and 48 years with an average of 36 years (SD = 10.45). Concerning the highest level of education, 11 students finished their Bachelor, 16 held a Master's degree and 3 a PhD. Participants' native language was Spanish, but they indicated (on 5pt.-rating scales) that they had good writing (M = 3.57, SD = 0.84) and reading (M = 4.13, SD = 0.72) skills in English. Participants were familiar with e-learning environments (M = 4.43, SD = 0.63) and slightly preferred online-courses over face-to-face courses (M = 3.57, SD = 0.94). Students gave their consent to participate in the study by filling out the first out of four questionnaires.

Due to their enrollment in the master degree program, all students had already acquired basic skills for online learning. The activities for this study were introduced by the course professor (teacher) to increase students' motivation for fulfilling the required tasks (Ko and Young, 2011).

3.3 Experimental Design

One main goal of the study was to test the generated questions quality with a balanced design covering the following aspects: The first factor concerns the question type and comprises the three factor levels multiple choice (MC), true or false (TF) and fill-in-the-blank (FiB) questions. Open ended questions were not included, because the automatic assessment cannot account for different wordings. The second factor refers to the creator of the concepts on which the questions are based on and has two levels, teacher and EAQC. Hence, the concept is either extracted by the teacher or automatically by the EAQC. The third independent variable refers to the creator of the questions. As the question is either generated by the teacher or by the EAQC, there are also the two factor levels, teacher and EAQC. Thus, the evaluation of the questions' quality is based on a 3 x 2 x 2 design. All questions were presented for two learning contexts, namely Problem-based learning and Project-based learning.

The dependent variables concern the quality and difficulty of the questions and their respective answers. The following evaluation criteria have already been applied by Hoefler et al. (2012) and go back to the work of Cannella et al. (2010). Quality of questions is measured by the four aspects pertinence, level, terminology, and difficulty. Pertinence denotes the relevancy of a question with respect to the topic. Level addresses whether a question is trivial or expresses a significant meaning. Terminology focuses on the appropriateness of the words chosen and difficulty refers to the perceived difficulty of a question. Quality of answers is measured for FiB and MC questions by means of the aspects terminology of the correct answer, ambiguity of the answer, and for MC questions also by the quality of the given distracters. For both, question and answer quality mean scores were calculated from the 4 respectively 2 (or 3) single aspects. All aspects were evaluated by the participants on 5-pt. rating scales with 1 indicating a low quality and 5 indicating a high quality (except for ambiguity, where 5 indicates high ambiguity and therefore low answer quality). According to the experimental design, we calculated a multivariate ANOVA with three factors.

Regarding the concepts extracted from the two learning contexts, we differentiated between teacher, EAQC, and student concepts. To evaluate the quality of these concepts participants rated their relevancy on a 5pt. rating scale.

3.4 Research Instruments and Procedure

During the self-regulated learning experiment, participants had to read two texts, extract concepts from these texts, take knowledge tests, and fill out four questionnaires. The questionnaires which were presented via Lime Survey, an open source survey application tool (see http://www.limesurvey.org/). It took five weeks to carry out the entire study starting with the selection of learning material until the presentation of the last questionnaire.

The two texts were provided by the teacher of the course and concerned the topics “Problem-based learning” and “Project-based learning”. They had 1307 and 1002 words respectively and dealt with basic knowledge (definition, history, theoretical foundations, etc.) on the two topics.

The experiment consisted of four phases (Phases 1 – 4). The first phase took place before the students started working on IWT. In this pre-phase students were asked to fill in a pre-questionnaire (Q1), which covered the following sections: demographic data, previous knowledge regarding e-learning, general questions about learning preferences, evaluation of question types, and students’ English skills. Furthermore, in Phase 1 the two learning resources were selected by the teacher. For both texts the EAQC and the teacher extracted 10 concepts each and put them in an order of relevance. For each topic, the extracted concepts were collected in an evaluation questionnaire (Q2), which was given to the students in Phase 2 in order to evaluate the concepts’ relevancy. In the case of equivalent concepts, the next one in the order of relevancy was chosen. The 20 concepts in each questionnaire were randomized, i.e. students did not know which concepts were extracted by the EAQC and which ones by the teacher. Hence, we could check whether the quality of the EAQC concepts is as good as the quality of the teacher’s concepts.

For the second phase students were assigned to two groups (Group 1 and Group 2). First, each group was asked to read one learning resource presented via IWT. Group 1 read the text on Problem-based learning, Group 2 the text on Project-based learning. Second, the students, in this phase, had to extract at least 6 concepts from the text by tagging and highlighting keywords. Additionally they were asked to put their concepts in an order of relevance. At the end of Phase 2 students answered the evaluation questionnaire Q2 with the teacher and EAQC concepts concerning the learning resource they had just read before.

In the third phase of the study questions based on the 20 concepts extracted in Phase 1 were generated by the teacher and the EAQC as follows: For each learning resource, the teacher was asked to generate one question for each of the six most relevant concepts extracted and ordered by herself and by the EAQC, four teachers’ and EAQC concepts were discarded. The only constraint was to use each questions type (MC, TF, FiB) twice for the teacher as well as for the EAQC concepts. Thereafter, the EAQC questions were generated for the same 12 concepts under consideration of the question type chosen by the teacher. Thus, we obtained two parallel questions for each concept. This approach was taken to make sure that for each concept a suitable question type was chosen and to ensure a fair comparison of teacher and EAQC questions.

Summarized, for each learning resource, 24 questions (12 teacher / 12 EAQC questions) based on 12 concepts (6 teacher / 6 EAQC concepts) were generated. This resulted in four different variants of questions: teacher questions based on teacher

concepts, teacher questions based on EAQC concepts, EAQC questions based on teacher concepts, and EAQC questions based on EAQC concepts. Combining the four question variants with the three types (MC, TF, FiB) yielded 12 different sorts of questions. From these questions the evaluation sections of Q3 and the knowledge tests for Phase 4 were constructed. For each topic, two parallel test forms were created and students were assigned to 4 groups (Group A, B, C and D). Students from Group 1 were assigned to Groups A and B, Group 2 was divided into Group C and D. Each pair of groups received parallel test and evaluation forms. Group A for example received a teacher question based on the first concept and an EAQC question based on the second, whereas Group B received those backwards, and so on.

Phase 4 was the second unit with students working in the IWT. First, students learned the text on IWT, which they had read in Phase 2. Second, they were asked to extract six concepts as in Phase 2 in order to compare consistency of the extracted concepts. Then, the students should generate "on-the-fly" questions from the EAQC based on their self-extracted concepts. After that, students received the knowledge tests created in Phase 3 about the text, they had just learned before. Groups A and B received the parallel tests on Problem-based learning, and Groups C and D the tests on Project-based learning.

After this, the students switched topics and read through the other learning resource. Groups A and B read the text on Project-based learning, and Groups C and D the one on Problem-based learning. Then, the students received Q3, in which they were asked to evaluate (a) the tool's usability, (b) the 10 most frequent student concepts extracted by their peers in Phase 2, and (c) the 12 questions generated by the teacher and the EAQC. Groups A and B evaluated the concepts and questions regarding the text Project-based learning, Groups C and D the concepts and questions related to Problem-based learning. Thus, Groups A and B evaluated the concepts extracted by students which were in Group C or D and the questions which Groups C and D had received in their knowledge tests and vice versa.

Finally, after students had finished their tasks in Phase 4, a post-questionnaire (Q4) was sent out concerning their motivation during the study. It contained the subscale "Task Value" from the MSLQ by Pintrich et al. (1991). This scale measures students' perception of the course material in terms of interest, importance, and utility. High task value should lead to more involvement in one's learning outcome and Pintrich et al. (1991) found a high correlation between task value and intrinsic goal orientation (r = .68). More specifically, students have to indicate their (dis)agreement to six questions regarding the task value. Furthermore, Q4 contained questions regarding students' motivation to do the different tasks involved in the study, e.g. reading the texts, extracting concepts, or working with the IWT. Answers were given on a 5pt. rating scale ranging from (1) not motivated at all to (5) very motivated.

Additionally, a post-questionnaire for the teacher was provided, which included questions on the usability of IWT, emotional aspects, and some open questions. We used the SUS (System Usability Scale) by Brooke (1996) in order to investigate the tool's usability. For the emotional status of the participant we added a scale (Computer Emotion Scale) by Kay and Loverock (2008) developed to measure emotions related to learning new computer software, describing four emotions: Happiness, Sadness, Anxiety and Anger.

4 RESULTS

From the 30 participants filling out the pre-questionnaire (Q1), 25 took part in Phase 2, in which they evaluated the relevancy of 20 concepts extracted by the teacher and EAQC. In Phase 4, we collected data from 20 participants, who all evaluated 10 student concepts and the 12 questions created by the teacher and EAQC. The knowledge tests which also contained the EAQC and teacher questions were taken by only 13 students. The data of one student was not included in the analysis below, because she called and saved the test, but did not answer a single question. Thus, for each of the two topics, six students took the prepared knowledge test. Furthermore, four students took an on-the-fly test with a total of 21 generated questions (seven for each question type).

The two texts were not chosen by the experimenters, but by the teacher of the course herself. Independent samples t-tests performed for concept relevancy (RelConc), mean question quality (QualQu), and mean answer quality (QualAns) yielded no differences between the two topics Problem-based and Project-based learning (RelConc:$t_{(68)}$=.155, p=.877; QualQu:$t_{(238)}$=.715, p=.475; QualAns:$t_{(158)}$=.243, p=.808). Thus, for the following statistical analysis the data were aggregated across the two topics.

4.1 Relevancy of Extracted Concepts

One important requirement for the generation of high quality questions is the extraction of relevant concepts within a given context. Thus, in Phase 2 of the experiment (see Section 3.4) the 10 most relevant concepts extracted by teacher and EAQC were given to the students to be evaluated (Q2). Additionally, 10 concepts extracted by the students in Phase 1 were presented to the other half of students in Phase 4 (Q3) and also evaluated with regard to relevancy. Thus for the comparison of EAQC and teacher concepts paired samples are available, whereas comparisons with student concepts involve independent samples. Figure 1 shows the average ratings for the three concept types for each topic. Across the topics mean ratings, teacher concepts, from 25 (for teacher and EAQC concepts) and 20 (for students) had the best rating over 4.12 (SD = .44).

A one-way ANOVA for the three concept extractors showed a significant effect ($F_{(2,67)} = 3.66$, $p = .031$). Post-hoc tests after Scheffé's method revealed a difference between teacher and student concepts ($p = .032$) but not between EAQC and teacher or student concepts ($p = .353$ and .42 respectively). Since teacher and EAQC concept evaluations are based on the same sample, we also performed a repeated measures t-test for a stricter comparison of these ratings; however, with $t_{(24)} = 2.0$ and $p = .057$ EAQC concepts do still not differ significantly from the concepts extracted by the teacher. Thus, it can be stated that concepts extracted from the EAQC are as relevant as concepts extracted by humans.

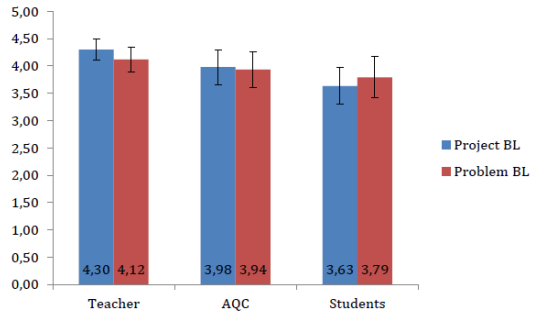

Figure 1: Relevancy ratings for concepts extracted by EAQC, teacher, and students.

4.2 Quality of Questions

To evaluate the quality of questions 3x2x2 MANOVAs were performed by aggregating the data from both topics (Problem-based and Project-based learning).

With 20 students performing the evaluation and 12 questions per student, 240 answers were collected for each criterion. These are divided equally over the question types (80 data points for TF, MC and Fib questions each), concept extractors (120 from teacher and EAQC each), as well as question creators (also 120 from teacher and EAQC each).

The evaluation metrics consisted of four measures for the quality of the questions themselves (pertinence, level, terminology, and difficulty) and two and three measures for the quality of the answers of FiB and MC questions (terminology of answer and ambiguity of answer for both, plus quality of distractors for MC). Since TF questions are not included in this analysis, the number of data points for answer quality decreases to 160 (80 for distractor quality). Figure 2 shows the mean ratings for question and answers quality per question variant (concept extractor x question creator) and question type (TF, MC, FiB). Because of the different numbers of questions types (TF, MC, Fib) involved, we performed two 3x2x2 ANOVAS for mean questions and mean answer quality.

The results are summarized in Table 1. Whereas none of the three factors had an effect on mean answer quality, we found one significant effect on mean question quality. More specifically, with $M_{teacher} = 3.57$ and $M_{AQC} = 3.32$ (SE = .063) questions created by the teacher were evaluated significantly higher, than those created by the EAQC. However, with a partial eta² value of .034 the effect size is rather small. Interactions are also non-significant. To investigate, in which aspect the questions differ, we had a closer look at the different aspects contributing to question (and answer) quality.

A three-way MANOVA including the four aspects of question quality yielded the expected effect of question creator for the multivariate results ($F_{(4,225)} = 2.51$, $p = .043$, $\eta_p^2 = .043$) and no other main effects or interactions. Univariate results showed that the effect is due to higher evaluations of teacher questions' pertinence and level. For pertinence teachers questions reached a mean rating of M = 3.95 as compared to M = 3.58 for EAQC questions with SE = .092 ($F_{(1,228)} = 7.82$, $p = .006$, $\eta_p^2 = .033$).

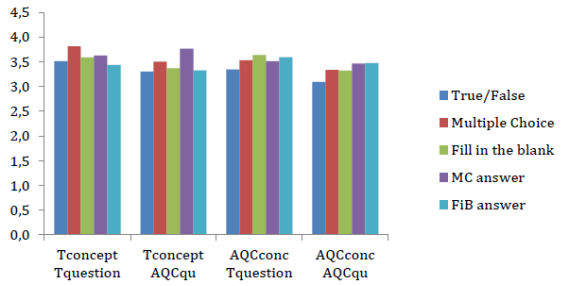

Figure 2: Mean rating for question and answer quality of T/F, MC, and FiB question (Tconcept = concept extracted by teacher, EAQCqu = question created by EAQC, etc.).

Table 1: Effects from three-way ANOVAS on mean question and answer quality.

	Factor	*df*	*F*	*p*	η_p^2
Mean question quality (*N* = 240)	Question Type	2,228	2.432	.09	.021
	Concept extractor	1,228	2.399	.123	.01
	Questions creator	1,228	8.025	.005	.034
Mean answer quality (*N* =160)	Question type	1,152	1.084	.299	.007
	Concept extractor	1,152	.044	.834	.000
	Question creator	1,152	.070	.792	.000

Ratings for level imply that teacher questions are more meaningful (M = 3.62) than EAQC questions (M = 3.27, SE = .096), with $F_{(1,228)} = 6.56$, p = .011, $\eta_p^2 = .028$. However, for both aspects the effects are only small in size. Ratings for questions' terminology and perceived difficulty did not differ significantly, and there were no significant interactions among the three factors.

With respect to the quality of answers, a closer look at the three aspects revealed no significant effects and no interactions for the multivariate results of the performed MANOVA (for MC and FiB questions), for the aspect terminology of the answer, and for the quality of distracters (ANOVA for only MC questions). However, we did find an effect of question type on the ambiguity of answers. With M = 2.35 (SD = 1.17) for MC and M = 2.8 (SD = 1.26) for FiB questions, participants evaluated answers of the latter question type to be more ambiguous.

The results presented in this section clearly show that the answers to the questions provided from the tool are relevant. Moreover, participants evaluated the quality of answers generated by the EAQC as equally high as the quality of answers generated by the teacher. By using teacher concepts for half of the questions created by the EAQC, we could also show that the tool is able to generate questions from concepts entered by users. Regarding the question whether all types of questions generated from the tool are as high in quality as questions generated by humans, the answer is two-fold. Whereas the quality of answers, the terminology, and the perceived difficulty of EAQC questions are evaluated equally high as those of teacher questions, the level and pertinence of questions received lower ratings. Thus, teacher questions seem to be less trivial and address the topic in a more meaningful way.

4.3 Difficulty of Questions

To further investigate if all types of questions generated from the tool are as high in quality as questions generated by humans, we also collected data concerning the real difficulty of questions. Therefore, the same questions which were presented for evaluation were also prepared as knowledge test and uploaded to the courses in the IWT. To avoid an influence of the evaluation process on the test taking or vice versa, each test was given to half of the students as test and to the other half for evaluation purposes. Whereas 20 students did the evaluation of questions only 13 took the knowledge test. Data of 12 students who each answered 12 questions could be analysed. All together 45.83% of the questions were answered correctly, which equals 66 out of 144 questions. Table 2 gives an overview on how many items per questions variant have been answered correctly. Since there is no difference between the topics Problem-based and Project-based learning (32 vs. 34 correct responses), the data are aggregated across the two topics. We calculated χ^2 tests to compare the frequencies for questions that (a) are based on EAQC vs. Teacher concepts, (b) are generated by EAQC or teacher, and (c) are designed as either TF, MC, or FiB question. Except for the question type, the critical χ^2 values exceeded the empirical ones. With $\chi^2 = 22.46$, the differences between the three question types are statistical significant, which can clearly be attributed to the very low solution rate for fill-in-the blank questions (8% correct solutions compared to 69% for MC and 60% for TF).

To investigate the relationship between perceived difficulty of actual difficulty, we correlated the mean ratings and number of correctly solved items per questions variant (i.e. for 12 different item types, as e.g. TF with teacher concept and teacher questions or MC with EAQC concept and teacher questions, etc.). The resulting correlation of $r_{(12)} = -0.71$ ($p = .009$) indicates that questions which are perceived as more difficult are also solved

by less participants.

We also looked at the difficulty of the on-the-fly questions. From seven questions per type, five TF, six MC and 1 FiB have been answered correctly, which is in line with the findings reported above (high difficulty of FiB, no difference between TF and MC).

Table 2: Number of correctly answered questions per question type.

	Teacher concept		AQC concept		
	Teacher questions	AQC question	Teacher question	AQC questions	
TF	9	6	11	3	29
MC	0	15	6	12	33
FiB	2	0	0	2	4
Sum	11	21	17	17	66
Sum concept	-	32	-	34	
Sum question	-	-	28	38	

4.4 Usability of the EAQC Integrated in IWT

The basis for the validity of usability measures is the time participants spent in the system. Log data show that students accessed the IWT on average 3.31 times (SD = .72, MIN = 1, MAX = 7) and spent M = 103.78 min (SD = 89.3) within the system. The teacher (and her assistant) spent together 39 hours 28 minutes in the IWT, accessing it 102 times.

To evaluate the usability of the integrated EAQC, the SUS scale (see Section 3.4) was presented to students as well as the teacher. Students' mean SUS scores amount to 57.59 (SD = 16.99) with a rather large range from 23.75 up to 87.5. The SUS score provided by the teacher was 48.13. Thus students perceived the usability of the EAQC integrated in the IWT very differently, but on average higher than the teacher. However, both student and teacher scores are below the average of 68, which is the reference value suggested by Brooke (1996). Despite the low SUS score and the great amount of time the teacher spent in the IWT, ratings from the Computer Emotion Scale (see Section 3.4) show very positive emotions while working with the system (mean scores for happiness/sadness/anxiety/anger in the given order are 3/1/1.25/1).

4.5 Motivational Aspects and Task Value

To evaluate whether the tool had a positive impact on users' motivation concerning their learning, an analysis with results from 14 students filling out the post-questionnaire is presented. A requirement for having a positive impact on students' motivation is that they are generally comfortable with SRL settings. Participants of this study indicated that they like SRL environments (M = 3.77, SD = .76) and that they prefer learning on their own over being supervised all the time (M = 3.8, SD = 1.01). Furthermore, they agreed on the statements that testing themselves helps when they learn something (M = 3.87, SD = .72) and that they need clear instructions when they learn something (M= 3.93, SD = 1.03). These results are in line with the comments on the tool itself, namely that they liked the EAQC and automatic assessment as well as the possibility to highlight and save important concepts to support their learning process.

The task value scale by Pintrich et al. (1991), which was presented in the post-questionnaire showed that students were highly interested in the task and also perceived it as being important and useful. Mean ratings to the six single questions ranged between 4.5 (SD = .65) and 4.71 (SD = .47) resulting in a mean task value of 4.58 (SD = .48). Due to the high correlation of task value and intrinsic goal orientation reported above, it can be assumed, that students were also intrinsically motivated and involved in their learning activities. This result is also supported by the high motivation ratings for the single tasks required during the study, which ranged between 3.77 (SD = 1.01) and 4.31 (SD = .85). Figure 3 shows the mean ratings for task value and motivation for doing different task.

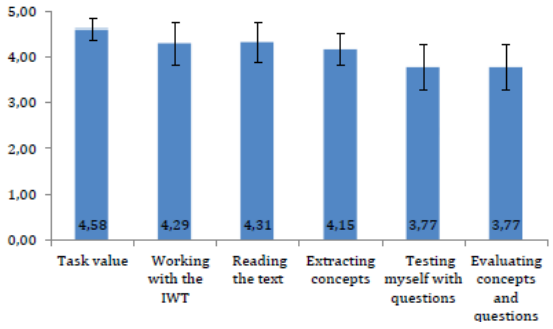

Figure 3: Mean task value (MSLQ) and level of motivation for various tasks.

4.6 Support of Self-regulated Learning

To investigate the pedagogical and psychological impact of the tool, we checked, whether the tool supports self-regulated learning and students can thus benefit from using the tool. According to the teacher the tool constitutes a support for students in the self-study process. From a teacher's point of

view, the functions tested by the teacher were too few to judge the worth of the tool for teachers.

From student's point of view, testing themselves with questions had a positive impact on their learning activities (M = 4.43, SD = .76), taking the course improved their understanding of domain concepts (M = 4.5, SD = .65), and the course was a worthy educational resource (M = 4.43, SD = .76). In their open comments, all 14 students stated that they would benefit from self-assessments (self-generated tests) when learning in general. More specifically, they said, that self-assessments are a good preparation for real assessments, that they help to know one's level of knowledge or progress, and that it helps to improve the understanding and retention of concepts. Only one student indicated that he wouldn't go through a self-generated test, because for him reflection on his knowledge is more important. Another student stated that the self-assessments help to study a text, but that the questions were not good.

5 CONCLUSIONS

The aim of the conducted study was to evaluate a tool for automatic question creation (EAQC) and its application within the IWT. All questions generated by the EAQC are based on concepts, which are in a first step automatically extracted from a given text. In a second step the EAQC creates for each concept the required types of questions (up to four different questions per concept). Thus, the quality of the extracted concepts is an important factor for the achieved quality of the generated questions.

In this study the teacher of an online course on e-learning provided two texts and extracted the 10 most relevant concepts out of each text. Simultaneously, the EAQC extracted concepts out the same texts and put them into an order of relevance (see Section 3.4). We compared the relevancy of concepts extracted by the EAQC, by the teacher, and by students. The obtained results show comparable quality of automatically and human (teacher and students) generated concepts. Thus, we can conclude that the tool is able to extract relevant concepts from a text, which form a suitable basis for knowledge questions.

The quality of questions and their respective answers was evaluated by comparing it to the quality of questions created by a teacher (both EAQC and teacher questions were based on an equal number of EAQC and teacher concepts). A three-way MANOVA including the factors questions type (TF, MC, FiB), concept extractor (teacher, EAQC), and question creator (teacher, EAQC) revealed an effect of question creator for the dependent variables pertinence and level. Thus, EAQC questions are equally well formulated as teacher questions (no effect on terminology) and are perceived as equally difficult, but they are evaluated as more trivial and less relevant than teacher questions.

However, considering that the factor "question creator" accounts for only about 3% of the overall (effect and error) variance (η_p^2 =.033 and .028 for the two measures) and that the EAQC is mainly meant as tool to support self-regulated learning, the outcome of the evaluation is definitely positive.

Results from the evaluation of answers revealed no difference between teacher and EAQC terminology, ambiguity, or distractor quality. The same is true for the actual difficulty of questions, indicated by the number of correctly solved items. Thus, the application of the EAQC in a real learning setting and the evaluation of the tool by postgraduate students yielded very promising results. In a next step, the evaluation process should also involve domain experts (e.g. a group of teachers) as well as experts in the field of assessment.

Regarding the tool's usability, SUS scores from both teacher and students were below average, but the teacher was still in a very positive emotional state and open comments from both sides show that they appreciate the tool and its functions. Students indicated that they would benefit from automatic self-assessments and that the course is a worthy educational resource. The results show high task values and high motivational ratings for the different tasks performed during the study. Thus, we can conclude that the EAQC and in general automatic question creators are able to motivate students in their learning activities and should be a fundamental part of SRL environments.

The results from the evaluation of questions generated automatically by EAQC in a broader setting are encouraging. The study, including participants with different cultural background in Latin America, allowed the researchers to test the tool and perception of the assessment in PLEs in order to look for worldwide solutions.

Besides the above mentioned evaluation studies with experts, future work needs to focus on improving the tool's usability, clarity, and performance. Also a focus on the quality of extracted concepts and questions with text in different languages should be considered.

ACKNOWLEDGEMENTS

This research was supported by the EC under the Project ALICE "Adaptive Learning via Intuitive/Interactive, Collaborative and Emotional Systems", Grant Agreement n.257639. We are very grateful to Isabella Pichlmair, Dominik Kowald, Marcello Rosciano, Patricia Lavin and GES Team for their great support in conducting this study.

REFERENCES

Agarwal, M., Shah, R., & Mannem, P. (2011). Automatic question generation using discourse cues. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*: 1–9.

AL-Smadi M., Hoefler M., & Guetl, C. (2011). Integrated and Enhanced e-Assessment Forms for Learning: Scenarios from ALICE Project. *Proceedings of Special Track on Computer-based Knowledge & Skill Assessment and Feedback in Learning Settings* ICL 2011, Piestany, Slovakia, 2011, pp. 626-631.

Bannert, M. (2006), Effects of Reflection Prompts when learning with Hypermedia, *Journal of Educational Computing Research*, vol. 35, no 4, pp. 359–375.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. *In Usability evaluation in industry*. London: Taylor & Francis.

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005) "Automatic Question Generation for Vocabulary Assessment", *Proc. of the Human Language Technology Conference on Empirical Methods in Natural Language Processing*: Canada, pp 819- 826.

Canella, S., Ciancimino, E., & Campos, M.L. (2010) "Mixed e-Assessment: an application of the student-generated question technique", Paper read at *IEEE International Conference EDUCON 2010*, Madrid.

Capuano, N., Gaeta, M., Marengo, A., Miranda, S.,Orciuoli, F., & Ritrovato, P. (2009). LIA: an Intelligent Advisor for e-Learning. *InteractiveLearning Environments,* Taylor & Francis, vol. 17, no. 3, pp. 221-239.

Chen, W., Aist, G., & Mostow J. (2009).Generating questions automatically from informational text. *In Proceedings of the 2nd Workshop on Question Generation*: 17- 24. UK: Brighton.

Conati C., & Vanlehn K. (2000), "Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation "*International Journal of Artificial Intelligence in Education*, vol. 11, pp. 389–415, 2000.

Dochy, F. J. R. C., & McDowell, L. (1997). "Assessment as a tool for learning." *Studies in Educational Evaluation*, 23 (4), pp. 279-298.

Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). "Automatic Generation System of Multiple-choice Cloze Questions and its Evaluation", *Knowledge Management & E-Learning: An International Journal (KM&EL),* Vol 2, No 3, 2010.

Guetl, C., Lankmayr, K., Weinhofer, J., & Hoefler, M.(2011). Enhanced Automatic Questions Creator – EAQC: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education. *Electronic Journal of e-Learning*, 9, 23-38.

Hoefler, M., AL-Smadi, M., & Guetl, C. (2011). Investigating content quality of automatically and manually generated questions to support self-directed learning. In D. Whitelock, W. Warburton, G. Wills, and L. Gilbert (Eds.).*CAA 2011 International Conference*, University of Southampton.

Hoefler M., AL-Smadi M., & Guetl C. (2012). Investigating the suitability of automatically generated test items for real tests. *The International Journal of eAssessment,* 2(1) [Online]. Available: Doc. No.26.http://journals.sfu.ca/ijea/index.php/journal/-article/viewFile/27/26.

Kay, R. H., & Loverock, S. (2008). Assessing emotions related to learning new software: The computer emotion scale. *Computers in Human Behavior*. 24, 1605-1623.

Ko, C., & Young, S. (2011). Explore the Next Generation of Cloud-Based E-Learning Environment. *Edutainment Technologies. EG LNCS*, Vol. 6872, 2011, pp 107-114

Kroop S., Berthold M, Nussbaumer A., & Albertet D., (2012). "Supporting Self-Regulated Learning in Personalised Learning Environments" *1st Intl Workshop on Cloud Education Environments*, CEUR Vol. 945 ISSN 1613-0073, 2012.

Liu, M., & Calvo, R.A. (2009). An automatic question generation tool for supporting sourcing and integration in students essays. *14th Australasian Document Computing Symposium*, Sydney, Australia.

Myller, N. (2007). Automatic generation of prediction questions during program visualization. *Electronic Notes in Theoretical Computer Science* 178: 43–49.

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),198–218.

Pintrich, P.R., Smith, D.A.F., Garcia, T., & McKeachie, W.J. (1991). A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). Technical Report, 91, 7-17.

Stanescu, L., Spahiu, C. S., Ion, A., & Spahiu, A. (2008).Question generation for learning evaluation. *Proceedings of the International Multiconference on Computer Science and Information Tech*, 509-513.

Villalon, J. and Calvo, R. A. (2009) "Concept Extraction from student essays, towards Concept Map Mining", *Proceedings of the 9th IEEE International Conference on Advanced Learning Technologie*, pp 221-225.

Zimmerman, B. J. (1989). 'Models of self-regulated learning and academic achievement.' In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theory, research and practice*. New York: Springer-Verlag.