2010 10th IEEE International Conference on Advanced Learning Technologies

Towards an Enhanced Approach for Peer-Assessment Activities

Mohammad AL-Smadi
Graz University of Technology
Graz, Austria
msmadi@iicm.edu

Christian Gütl
Graz University of Technology,
Austria, Curtin University of
Technology, Perth, WA.
cguetl@iicm.edu

Frank Kappe
Graz University of Technology
Graz, Austria
Frank.kappe@iicm.edu

Abstract— **In this paper, we will address an enhanced approach for online peer-assessment where new features of candidate answer marking have been used. Students are capable to select specific parts from the candidate answer and mark them as *correct, wrong, or irrelevant.* Special colors are used to tag the selected part of the candidate answer in order to help students giving a reasonable final score and to provide visual feedback for the answer owner. A web based tool has been developed, an experiment was conducted and valuable results have been found.**

Keywords- peer, self-assessment; computer-assisted/based assessment; e-learning.

I. INTRODUCTION

Peer-assessment has been defined as "an arrangement for the peers to consider the level, value, worth, quality or successfulness of the products or outcomes of learning of others of similar status" [1]. From this definition, you can notice that peer-assessment is not a method for measurement but it is a source of assessment that can be utilized within a framework side by side with other assessment methods [2]. Peer-assessment has gained more importance from its emphasis on the necessity of making the student an important part of the assessment process not only as assessee but also as assessor where students and tutors collaboratively work together in the assessment model [3]. Rather than supporting the learner-centered model, peer-assessment may decrease staff load and time consumed on the assessment process as well as it may develop certain skills for the students such as, communication skills, self-evaluation skills, observation skills and self-criticism [4].

Several tools have emerged since the beginning of the 21st century. Some of these tools are part of computer-based assessment systems that implement peer-assessment methods [5]. The earliest reported system to support peer-assessment developed at the University of Portsmouth, "*The software provided organizational and record-keeping functions, randomly allocating students to peer assessors, allowing peer assessors and instructors to enter grades, integrating peer- and staff-assessed grades, and generating feedback for students*" [6]. One of the first systems with the peer-assessment methods was a tool for collaborative learning and nursing education based on multi-user database, which was called MUCH (Many Using and Creating Hypermedia) [7]. In the late 1990s, NetPeas (Network Peer Assessment System) has been implemented, and Artificial Intelligence (AI) has been used to develop the tool of Peer ISM that combines human reviewing with artificial ones [7]. Computer-assisted-peer-assessment systems have been also affected by the revolution of World Wide Web (WWW). An example of the first reported web-based system was a web-based tool for collaborative hypertext authoring and assessment via e-mail [8]. Other systems such as, a web-based system for group contributions on engineering design projects [9], the Calibrated Peer Review (CPR) which was introduced in 1999 [10], the Peer Grader (PG) as a web-based peer evaluation system [11], The Self and Peer Assessment Resource Kit (SPARK) which is an open-source system designed to facilitate the self and peer assessment of groups [12], The computerized Assessment by Peers (CAP) is another example [5]. Further examples such as, OASIS which has automated handling for multiple-choice answers and peer assessment for free-text answers, The Online Peer Assessment System (OPAS), which has some abilities for assignment uploading and reviewing as well as groups management and discussions [13], An improvement for this system was introduced in Web-based Self and Peer Assessment (Web-SPA) system to avoid the lack in determining standards, methods of scoring and the workflow of the assessment process [14]. Recent examples of peer-assessment developments are, the enhanced open-source implementation of WebPA system which was originally developed in 1998 [15], as well as the Comprehensive Assessment of Team Member Effectiveness (CATME) system which assesses the effectiveness of team members contributions [16].

This paper focuses on how a computer-assisted peer assessment can motivate students to participate in the learning process as well as to provide them with added value. Moreover, discussing the following aspects related to peer assessment: *(A1) Reliability of peer-assessment results, (A2) Appropriate measurement for peer-assessment performance, (A3) Motivation and attitudes, (A4) Knowledge acquisition, And (A5) Usability aspects.* Therefore, a web-based prototype has been developed and an experiment has been conducted. To this end, the rest of this paper is structured as follows: Section 2 describes peer assessment procedure and the experiment setup. Section 3 discusses the experiment results and Section 4 concludes this paper.

II. EXPERIMENT PROCEDURE

The experiment was performed as an e-learning activity for the course of "Information Search & Retrieval (ISR)" at Graz University of Technology in the winter term

978-0-7695-4055-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICALT.2010.181

2009/2010. The experiment was conducted in a controlled environment in the computer lab with a supervision of the course lecturer. A web-based peer-assessment tool was used by the students to participate in the experiment which is also used by the tutors in the evaluation process of the students' candidate answers. The experiment details are as follows:

- Introductory talk (10 minutes): at the beginning of the experiment a short introduction was given by the ISR course lecturer about the subject domain as well as the assessment in general and the peer-assessment as an emerging form of assessment. The importance of knowledge acquisition and knowledge assessment in modern learning settings was discussed briefly. The learning objectives behind this experiment were mentioned. The lecturer also stressed on the importance of the students performance during the experiment and clarified that the performance will be given 10 points as part of the overall grade for both the online test and the online peer assessment session of 5 points each.
- Online learning session (45 minutes): "Document Classification" as one of the main topics of ISR course was chosen to formulate the online learning material of the experiment. The material language is English and it has been extracted from Wikipedia. The material is formulated out of four web-pages and an introduction one, where the students were allowed to access and navigate between them as well as a set of further readings hyperlinks related to the subject domain.
- Online testing session (15 minutes): The knowledge that was gained by the student from the last session is assessed in this session. An English test language of five questions was deployed for the students as a web-based assessment activity. During this session the students were not allowed to access any course materials. The test items were variable, where the first questions was a definition one, the second was an enumeration, the third and the fourth were asking for a concept explanation while the fifth was an abbreviation. For each of the fifth questions a short-free answer and a confidence value out of 5 (i.e. "0" as very poor to "5" as very good) had to be provided. The confidence value is used to evaluate the level of maturity for the student answer (self-assessment).
- Break (15 minutes).
- Online reference answers preparation (15 minutes): During this session, the students were asked to prepare reference answers for the questions 1, 2 and 5. Differently from the last session the students were asked to access the course content and other useful materials to help them in preparing the reference answers.
- Online peer assessment session (45 minutes): in this session the students used the reference answers from the last session to evaluate and to peer-assess their answers from the online test session. Every student had to evaluate around 21 randomly selected answers for 3 different questions as well as 15 pre-prepared calibrated answers. For each answer, the students were capable to select parts from the candidate answer and mark them as *correct, wrong, or irrelevant.* Special colors are used to mark the selected part of the candidate answer based on its correctness (i.e. correct as green, wrong as red and irrelevant as yellow). A score should also be provided by the student for the answer from "0" (very poor) to "5" (very good). Using colored marks for the candidate answer supports the students for scoring the answer and to provide a reasonable score based on his colored marks. Moreover, the colored marks will be provided as a valuable feedback to the student who wrote this answer. Input-boxes for missing parts of the answer and additional notes were provided for the students to write into them as in Fig. 1.
- Experiment questionnaire (10 minutes): the students were asked to fill in a questionnaire that diagnoses their attitudes about the assessment activity of its three parts self-directed, online test and the peer-assessment one, as well as the usability of the web-based assessment prototype and their suggestions for further enhancements and notes.
- Results delivery: the students peers marks of their candidate answers have been used to compute the online test performance grade and provided as feedback as in Fig. 2.

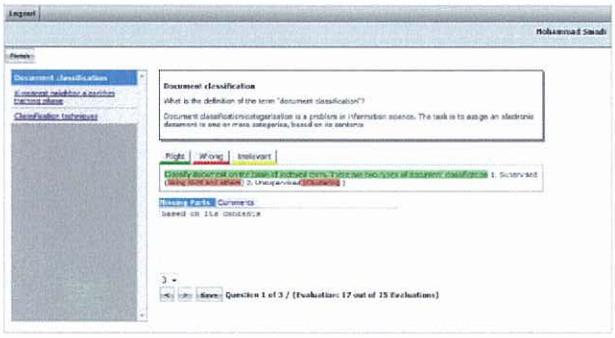

Figure 1. Screenshot from the peer-assessment step.

In order to compare the students' peer-assessment results with a reference grading values, a set of tutors had participated in the experiment. The tutors' peer-assessment process was as follows:

- Experiment Introduction: an e-mail was sent to all the tutors, in which a brief introduction about the experiment goals and procedures were outlined.
- Reference answer preparation: the tutors were asked to use the course content and other related materials to prepare a set o f reference answers

that they will use later on in the evaluation process.

- Online peer-assessment: in this step, all the candidate answers from the students were evaluated by the tutors. The same colored marking facilities of some parts of the candidate answers were used. As well as the possibility of adding notes and missing parts of the candidate answers.

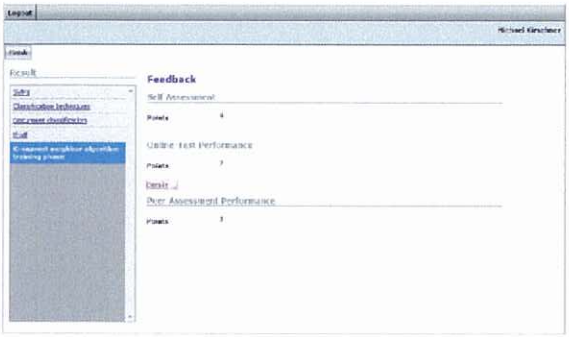

Figure 2. Screenshot from the feedback step.

A group of 25 students enrolled at the course of ISR. All of them participated in the experiment. 13 (52%) of the students were taking part in the course as a bachelor program, where 10 (40%) were master students, and 2 (8%) were doctoral students. 3 (12%) were females and 22 (88%) were males. The average age of the students was 26.7 years old with a minimum age of 22 and a maximum one of 36. The tutors were a group of 5 PhD students at the IICM (Institute for Information Technology and Computer Media) of Graz university of Technology. All of them were males and have a master degree of computer science.

III. RESULTS ANALYSIS

A. Students Questionnaire

As part of the students phase, they have been asked to fill in a questionnaire regarding their attitudes and comments on the experiment. The questionnaire diagnosed student's knowledge acquisition, learning attitudes, and the usability of the tool.

The self estimation of students' knowledge acquisition has been discussed in several researches [17]; [18]; [19]. Fig. 3 shows the results for the students' self estimation of knowledge acquisition from the overall experiment. From the students' point of view, their basic knowledge in the subject before the experiment was with a mean value of 3.84 (σ = 1.31) where ("0" represents fully disagreement and "5" represents fully agreement). The knowledge gained from the online learning phase was with a mean value of 4.12 (σ = 1.2). Preparation of reference answers has supported the students to get better knowledge in the subject domain with a mean value of 4.64 (σ = 1.04), where the knowledge that they had gained from the peer assessment task was with a mean value of 4.40 (σ = 1.35). Furthermore, students had used the course content during the peer assessment task with a mean value of 3.40 (σ = 1.80), where for them it was appropriate to use only the reference answers for evaluating the candidate answers with a mean value of 4.64 (σ = 1.15).

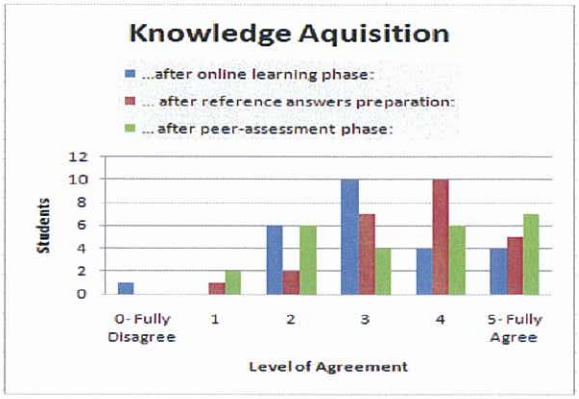

Figure 3. Student's self estimation of knowledge acquisition.

By analyzing the students' attitudes on the peer-assessment as part of modern learning settings, 15 (60%) students like to have peer-assessment as part of their future learning activities. Marking the candidate answers (right, wrong, and irrelevant) helped the students to better mark and score the answers with a mean value of 3.80 (σ = 1.47). The students argued that the time of the peer-assessment phase was too long with a mean value of 3.40 (σ = 1.38) where the suggested time for this phase was with an average value of 56.4 minutes (i.e. 11.4 extra minutes than the given time). They also argued that the required candidate answers per question to be evaluated were too many with a mean value of 4.08 (σ = 1.44) where they prefer the number per question to be with a mean value of 11.28 (σ = 6.65) candidate answer (i.e. 0.62 less than the required number which was 12 answers/question). 16 (64%) students think that it is a good idea to consider the quality of their peer-assessments as part of the final mark.

To get better idea about the usability of the tool, students were asked in the questionnaire about the tool functionalities and usability. According to the questionnaire, the overall usability of the tool was with a mean value of 3.96 (σ = 1.31) where ("0" represents fully unusable and "5" represents fully usable).

B. Tutors Phase

Because of the diversity in tutors experience the weighted mean has been chosen to compute the reference marks for the candidate answers. Table 1 shows the tutors experience represented in weights. The weights given to the tutors have been decided based on the tutors experience as well as the arithmetic mean of tutors grading from table 1 where a grade value of 2.5 represents the reasonable mean of a scale between "0" and "5". All of the tutors are PhD students in computer science (CS) some of them have

advanced knowledge in information retrieval (IR) as well as in assessment activities (AS).

Table 2 outlines the cross-correlations of the tutors' assessment results as well as the comparison with the weighted mean values of the candidate answers. For all of the test items the cross-correlation values vary between 0.507 (T1, T4) and 0.717 (T2, T3) by a mean value of 0.61 ($\sigma = 0.15$). For test Item 1 which asks for a definition, it has the best cross-correlation values between 0.483 (T2, T5) and 0.765 (T1, T2) by a mean value of 0.62 ($\sigma = 0.20$). Test Item 2 which asks for an enumeration, the cross-correlation values are between 0.324 (T1, T5) and 0.833 (T2, T3) by a mean value of 0.58 ($\sigma = 0.36$). For test Item 3 the cross-correlation values are the worst because it asks for an explanation of a concept which is more complex than definition and enumeration types, they are between 0.291 (T1, T4) and 0.656 (T2, T3) by a mean value of 0.47 ($\sigma = 0.26$). The same findings can be found in the literature where the variance between the tutor's correlation values depends on their experience as well as on the complexity of the assessment task [20]; [21]; [22].

In order to investigate the results, the absolute error of the tutors' individual score values is computed as the difference between the weighted average and the tutor score per candidate answer. As in table 3, the absolute error for all of the test items is between 1.12 ($\sigma = 0.83$) as worst result and 0.51 ($\sigma = 0.42$) as best result. For test item 1 the absolute error varies between 1.10 ($\sigma = 0.91$) as worst result and 0.48 ($\sigma = 0.52$) as the best one. The best case can be seen in test item 2 which reflects the simplicity of the assessment activity done by this item as an enumeration item where the absolute error is between 1.12 ($\sigma = 0.65$) and 0.43 ($\sigma = 0.30$). Test item 3 as the most complex item has not only lower cross-correlation but also higher absolute error values between 1.14 ($\sigma = 0.95$) and 0.61 ($\sigma = 0.40$). Moreover, all the best results are achieved by Tutor 2 which shows that the more experience the tutor the lower absolute errors she/he has.

C. *Students Phase*

In order to compare the student's peer-assessment performance with the tutor's reference scores, the arithmetic mean of peer's individual results per candidate answer has been used and the absolute error as the difference between the student's arithmetic mean and the tutor's reference marks has been computed. For all the three test items the arithmetic mean of absolute error is quite low with 0.60 ($\sigma = 0.48$). For the three test items individually, test item 1 the arithmetic mean of the absolute error is 0.62 ($\sigma = 0.41$). Test item 2 has the lowest value of 0.47 ($\sigma = 0.38$) since it is easier to score an enumerated question than scoring a short-free answer. Test item 3 has a higher value with 0.72 ($\sigma = 0.61$) which reflects the complexity of the assessment activity done by this item as a concept explanation one. The correlation between the arithmetic mean of the student's evaluations and the tutor's reference marks for each candidate answer is quite strong with 0.84 for all the three test items, 0.88 for test item 1; 0.78 for test item 2; and 0.82 for test item 3. Fig. 4 represents a scatter plot for the tutor's reference grading in comparison with the students peer assessments for the three test items sorted in ascending order by the tutor's reference grading values.

TABLE 1. TUTORS WEIGHTS BASED ON THEIR EXPERIENCES AND GRADING.

	Experience				Grading	
	CS	*IR*	*AS*	*Weight*	*Mean*	*σ*
T1	☑	☑	☑ ☑	2	2.53	1.58
T2	☑	☑ ☑	☑ ☑	3	2.81	1.59
T3	☑	☑	☑	1	2.99	1.86
T4	☑	☑	☑	1	3.63	1.75
T5	☑	☑	☑	1	2.85	1.78

TABLE 2. CROSS-CORRELATIONS FOR TUTORS' ASSESSMENT RESULTS.

		T1	T2	T3	T4	T5	WMW
All Test Items	T1	1.000	0.610	0.551	0.507	0.565	0.811
	T2		1.000	0.717	0.604	0.576	0.912
	T3			1.000	0.567	0.608	0.820
	T4				1.000	0.531	0.743
	T5					1.000	0.755
	WMW						1.000
Test Item 1	T1	1.000	0.765	0.741	0.664	0.727	0.938
	T2		1.000	0.687	0.560	0.483	0.892
	T3			1.000	0.552	0.590	0.829
	T4				1.000	0.612	0.754
	T5					1.000	0.759
	WMW						1.000
Test Item 2	T1	1.000	0.497	0.426	0.628	0.324	0.712
	T2		1.000	0.833	0.536	0.565	0.926
	T3			1.000	0.559	0.721	0.875
	T4				1.000	0.383	0.699
	T5					1.000	0.718
	WMW						1.000
Test Item 3	T1	1.000	0.496	0.578	0.291	0.552	0.776
	T2		1.000	0.656	0.390	0.652	0.882
	T3			1.000	0.314	0.573	0.793
	T4				1.000	0.496	0.569
	T5					1.000	0.810
	WMW						1.000

TABLE 3. THE ABSOLUTE ERRORS FOR TUTOR'S ASSESSMENT PERFORMANCE.

	All Test Items		Test Item 1		Test Item 2		Test Item 3	
	Mean	*σ*	*Mean*	*σ*	*Mean*	*σ*	*Mean*	*σ*
T1	0.78	0.61	0.52	0.47	1.05	0.57	0.76	0.68
T2	0.51	0.42	0.48	0.52	0.43	0.30	0.61	0.40
T3	0.87	0.63	0.87	0.79	0.84	0.34	0.89	0.65
T4	1.12	0.83	1.10	0.91	1.12	0.65	1.14	0.95
T5	0.88	0.77	0.97	0.92	0.85	0.67	0.82	0.71

IV. CONCLUSIONS

Regarding the reliability of the peer-assessment results (A1), the level of agreement between the student's peer evaluations and the tutor's reference grading varies according to the complexity of the assessment task (represented by the test items), the experience of the individuals, as well as the motivation and attitudes. Experiment results showed for students as well as for tutors the highest level of agreement was for simple assessment tasks such as definitions and enumeration answers, where the level of agreement was fair with more complex assessment

activities such as concept explanation answers. A weighted average has been used to enhance the tutor's assessment values as they have different levels of experience. The average of the absolute error between the tutors weighted average grades and the students average marks for each candidate answer has been used to evaluate the performance of the students in the peer-assessment task (A2). By focusing on motivation and attitudes aspects (A3), overall students argued that the peer assessment task is an interesting alternative and they have gained new knowledge form it. Moreover, students and tutors acquired assessment skills and more detailed knowledge about the subject domain (A4). By focusing on usability aspects (A5), students in general liked the experiment procedure and they provided us with comments that can be considered as rooms of future improvements.

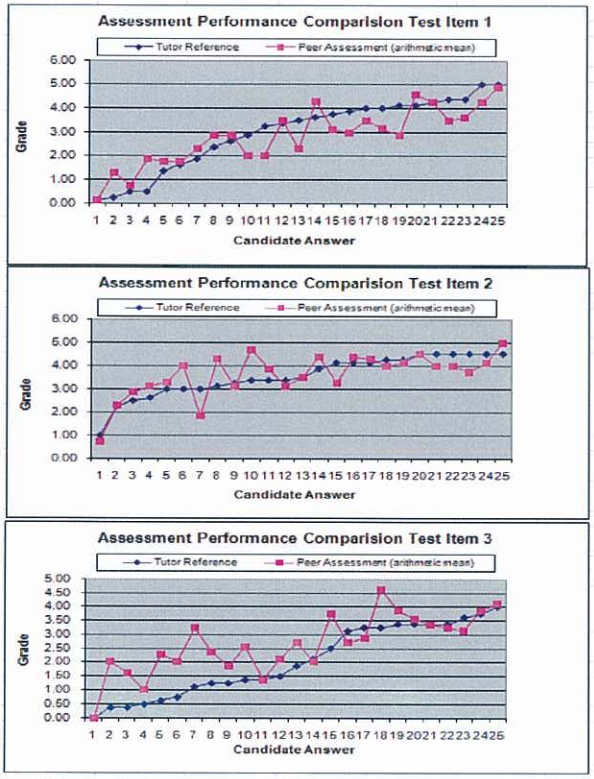

Figure 4. Students Peer-assessment Performance.

REFERENCES

[1] K. J. Topping; E. F. Smith; I. Swanson; A. Elliot, "Formative Peer Assessment of Academic Writing Between Postgraduate Students". *Assessment & Evaluation in Higher Education*, Vol. 25, No. 2, p. 150-169, 2000.

[2] G. Brown,; J. Bull, and M. Pendlebury. "What is assessment?". In *Assessing Student Learning in Higher Education*. London: Routledge, 1997.

[3] P. Orsmond, "Self- and peer-assessment: guidance on practice in the biosciences". In *Teaching Bioscience Enhancing Learning Series*, eds S. Maw, J. Wilson, and H. Sears, pp. 1-47 Leeds, UK: The Higher Education Academy Centre for Bioscience, 2004.

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

[5] P. Davies, "Peer-Assessment: No marks required just feedback? Evaluating the Quality of Computerized Peer-Feedback compared with Computerized Peer-Marking". *In Cook, J and McConnell, D (eds), Communities of Practice, Research Proceedings of the 10th Association for Learning Technology Conference (ALT-C 2003)*, 8-10, Sept 2003, Sheffield, UK.

[6] E. F. Gehringer, "Electronic peer review and peer grading in computer-science courses", Proc. of *the Technical Symposium on Computer Science Education*, p. 139-143, 2001.

[7] R. Rada,; S. Acquah,; B. Baker, & P. Ramsey,. "Collaborative Learning and the MUCH System". *Computers and Education, 20(3)*, 225-233, 1993.

[8] T. Downing,; & I. Brown,. "Learning by cooperative publishing on the World-Wide Web". *Active Learning 7*, 14-16, 1997.

[9] E. A. Eschenbach,; & M. A. Mesmer,. "Web-based forms for design team peer evaluations". *American Society for Engineering Education 1998, Annual Conference and Exposition*, Session 2630, 1998.

[10] L. Shepard,. "The Role of Assessment in a Learning Culture". *Educational Researcher, 29 (7)*, 4-14, 2000.

[11] E. F. Gehringer,. "Strategies and mechanisms for electronic peer review". In Proceedings, *Frontiers in Education Conference*, Vol 1., F1B/2 - F1B/7, 2000.

[12] M. Freeman,; & J. McKenzie,. "SPARK, a confidential web-based template for self and peer assessment of student teamwork: benefits of evaluating across different subjects". *British Journal of Educational Technology, 33(5)*, 551-569, 2002.

[13] S. Trahasch,. "Towards a flexible peer assessment system". *In Proceeding, Information Technology Based Higher Education and Training (ITHET 2004)*, 516-520, 2004.

[14] Y. T. Sung,; K. E. Chang,; S. K. Chiou,; & H. T. Hou,. "The design and application of a web-based self- and peer-assessment system". *Computer & Education, 45 (2)*, 187-202, 2005.

[15] WebPA, http://webpaproject.lboro.ac.uk. Last visited, 5th. February. 2009.

[16] M.W. Ohland, M.L. Loughry, R.L. Carter, L.G. Bullard, R.M. Felder, C.J. Finelli, R.A. Layton, and D.G. Schmucker, "The Comprehensive Assessment of Team Member Effectiveness (CATME): A New Peer Evaluation Instrument," *Proceedings of the 2006 ASEE Annual Conference*, Chicago, Illinois, June 2006.

[17] D. Magin,; & A. Churches,. "What do students learn from self and peer assessment?" *In Proceedings, EdTech'88 Conference, Australian Society for Educational Technology*, 27-29 September 1988.

[18] D.M.A. Sluijsmans,. "Student involvement in assessment.The training of peer assessment skills". *Unpublished doctoral dissertation, Open University of the Netherlands, The Netherlands.* 2002.

[19] P. McLaughlin,; & N. Simpson,. "Peer assessment in first year university: How the students feel". *Studies In Educational Evaluation, 30 (2)*, 135-149, 2004.

[20] Magin, D.; & Churches, A. (1988). "What do students learn from self and peer assessment?" In Proceedings, EdTech'88 Conference, Australian Society for Educational Technology, 27-29 September 1988.

[21] Sullivan, M; Hitchcock, M; & Dunnington, G.L. (1999). Peer and Self Assessment during Problem-Based Tutorials. *The American Journal of Surgery, 177 (March 1999)*, 266-269.

[22] Ward, M.; Gruppen, L.; & Regehr, G. (2002). Measuring Self-assessment: Current State of the Art. *Advances in Health Sciences Education, 7 (1)*, 63-80.