

Using Computers to Enhance Peer-assessment Activities: Towards a Flexible e-Assessment System

Mohammad AL-Smadi, Christian Gütl and Denis Helic

Institute for Information Systems and New Media (IICM), TU-Graz, Austria

Key words: *Peer-assessment, Self-assessment, On-line Testing, Computer-Assisted/Based peer-assessment, E-learning.*

Abstract:

Since the beginning of the 21st century, the learning process has changed from being repetitive to a new form of learning based on understanding, independency, learners' empowerment and skills improvement. As a main part of this learning process assessment is no more considered to discriminate between students, rather than it is used to enhance students learning and encourage them for further progress and success. In this new culture of assessment, teachers are no more considered to be knowledge carriers which they had to transfer to students' heads. Instead of that, students play major roles in assessment where new forms of assessment such as self-, and peer-assessment have been adapted. In this paper, we will address an enhanced approach of web-based peer-assessment for short free text answers. This approach is supposed to motivate students to participate in the assessment process, provide them with added value learning, as well as to maintain the reliability of the peer-assessment results. An experiment using this web-based system was conducted and valuable results were found.

1. Introduction

Peer-assessment is not new, it can be referred back to a long time of history where George Jardine the professor in the University of Glasgow from 1774 – 1826 prepared a pedagogical plan that included some peer-assessment methods and advantages [18]. 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” [19]. 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 [1]. Peer-assessment has gained its importance from its emphasis on the importance 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 [14]. 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 [5].

This paper focuses on aspects of peer assessment activities and how a computer-assisted approach can support as well as improve the assessment procedure and the learning process. A web-based prototype has been developed for implementing an enhanced peer assessment procedure and an experiment has been performed. To this end, the rest of this paper is structured as follows: Section 2 outlines related work for peer assessment activities. Section 3 describes the enhanced peer assessment procedure, and the experiment setup. Section 4 shows and discusses the experiment results and Section 5 outlines conclusions and future work.

2. Related Work

Several tools have been emerged since the beginning of the 21st century where some of them are computer-based assessment systems that implement the peer-assessment methods [4]. 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*” [10]. 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). In the same period a Macintosh application has been developed which has implemented a peer-review process for an assignment has been reviewed by two peers [9]; [10]; [15]. 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 [2]; [15]; [21]. Computer-assisted-peer-assessment systems has also affected by the revolution of World Wide Web (WWW), several web-based system have appeared later on. An example of the first reported web-based system was a web-based tool for collaborative hypertext authoring and assessment via e-mail [6]. Other systems such as, a web-based system for group contributions on engineering design projects [7], the Calibrated Peer Review (CPR) which was introduced in 1999 [3], the Peer Grader (PG) as a web-based peer evaluation system [9], The Self and Peer Assessment Resource Kit (SPARK) which is an open-source system designed to facilitate the self and peer assessment of groups [8], The computerized Assessment by Peers (CAP) is another example [4]. 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 [20], An improvement for this system was introduced in Web-based Self and Peer Assessment (Web-SPA) system to avoid the lake in determining standards, methods of scoring and the workflow of the assessment process [17]. Recent examples of peer-assessment developments are, the enhanced open-source implementation of WebPA system which was originally developed in 1998 [23], as well as the Comprehensive Assessment of Team Member Effectiveness (CATME) system which assesses the effectiveness of team members contributions [13].

3. Experiment Setup

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 2008/2009. The experiment was conducted in a controlled environment in the computer lab with a supervision of the course lecturer. A web-based Assessment system was used by the students to participate in the experiment which is also used by the tutors in the evaluation process of the students’ 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 domain of the subject 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 [11]. The material language is English and it has been extracted from Wikipedia [24]. 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 system. 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 10 had to be provided. The confidence value is used to evaluate the level of maturity for the student answer (self-directed 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 with a confidence value for their estimation of their answers quality. Differently from the last session the students were asked to access the course content and other useful materials to help them in identifying 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 30 randomly selected answers for questions 1, 2 and 5 as well as 15 pre-prepared optimal answers by the course teacher. For each answer, the students were asked to mark the answer by special tags for highlighting, underlining or changing to italic. Underlining some parts of the answer means that they are correct, where highlighting them means that they are wrong, and changing them to italic means that they are irrelevant. Input-boxes for missing parts of the answer and additional notes were provided for the students to write into them. A mark should also be provided by the student for the answer from “0” (very poor) to “10” (very good). Buttons were used to represent the candidate answers, they all yellow at the begging and once the student evaluates one of them its button color becomes green.
- Experiment questionnaire (10 minutes): the students were asked to fill in a questionnaire that diagnoses their impressions 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: as part of later on feedback provision the students’ answers and performance has been analyzed and a final grade has been sent to them by e-mail.

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 answers from the students (test and reference answers for the five questions) were evaluated by the tutors. The same marking and grading facilities of highlighting, underlining and changing to italics of some parts of the candidate answers were possible. As well as the possibility of adding notes and missing parts of the candidate answers.

A group of 27 students enrolled at the course of ISR. The students were separated into two groups 12 for the first group and 15 for the second. All of them participated in the experiment. 14 (51.9%) of the students were taking part in the course as a bachelor program, where 13 (48.1%) were master students. 3 (11.2%) were females and 24 (88.8%) were males. The average age of the students was 26.5 years old with a minimum age of 22 and a maximum one of 37. 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.

4. Results Analysis

4.1 Tutors Phase Results

With response to the diversity of experience that the tutors have, it has been decided to use the weighted mean instead of the arithmetic mean to compute the reference marks for the candidate answers. Table 1 shows the tutors experience represented in weights. All of the tutors are PhD students in computer science (CS) and doing well. Some of them have advanced knowledge in information retrieval (IR) as well as in assessment activities (AS). 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 5 represents the reasonable mean of a scale between 0 and 10. The cross correlations values from table 2 support that by noticing that the best correlations of assessment results were between (T1, T4) and (T2, T5) for the all test items.

Table 1: Tutors Weights based on their experiences and grading

	Experience				Grading	
	CS	IR	AS	Weight	Mean	σ
T1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	2	6.04	3.62
T2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	2	3.86	4.0
T3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1	7.31	3.39
T4	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	2	6.14	3.64
T5	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	3	4.21	3.69

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 0.499 (T2, T3) and 0.833 (T1, T4) by a mean value of 0.67 ($\sigma = 0.24$). For test Item 1 which asks for a definition, the cross-correlation values are between 0.555 (T1, T2) and 0.865 (T1, T4) by a mean value of 0.71 ($\sigma = 0.22$). A better situation for test Item 2 which asks for an enumeration, the cross-correlation values are between 0.701 (T3, T5) and 0.949 (T1, T2) by a mean value of 0.83 ($\sigma = 0.18$). For test Item 3 the cross-correlation values are the worst while it asks for an explanation of a concept, they are between 0.126 (T2, T3) and 0.755 (T1, T4) by a mean value of 0.44 ($\sigma = 0.44$). 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 [12]; [16]; [22].

Table 2: Cross-correlations for tutors' assessment results

		T1	T2	T3	T4	T5	WMW
All Test Items	T1	1.000	0.706	0.618	0.833	0.686	0.903
	T2		1.000	0.499	0.567	0.707	0.844
	T3			1.000	0.633	0.514	0.699
	T4				1.000	0.645	0.856
	T5					1.000	0.887
	WMW						1.000
Test Item 1	T1	1.000	0.555	0.764	0.865	0.645	0.891
	T2		1.000	0.716	0.613	0.642	0.788
	T3			1.000	0.826	0.675	0.893
	T4				1.000	0.771	0.941
	T5					1.000	0.857
	WMW						1.000
Test Item 2	T1	1.000	0.949	0.816	0.845	0.813	0.963
	T2		1.000	0.799	0.757	0.766	0.923
	T3			1.000	0.688	0.701	0.839
	T4				1.000	0.724	0.888
	T5					1.000	0.915
	WMW						1.000
Test Item 3	T1	1.000	0.392	0.303	0.755	0.640	0.866
	T2		1.000	0.126	0.311	0.534	0.554
	T3			1.000	0.318	0.180	0.353
	T4				1.000	0.570	0.836
	T5					1.000	0.896
	WMW						1.000

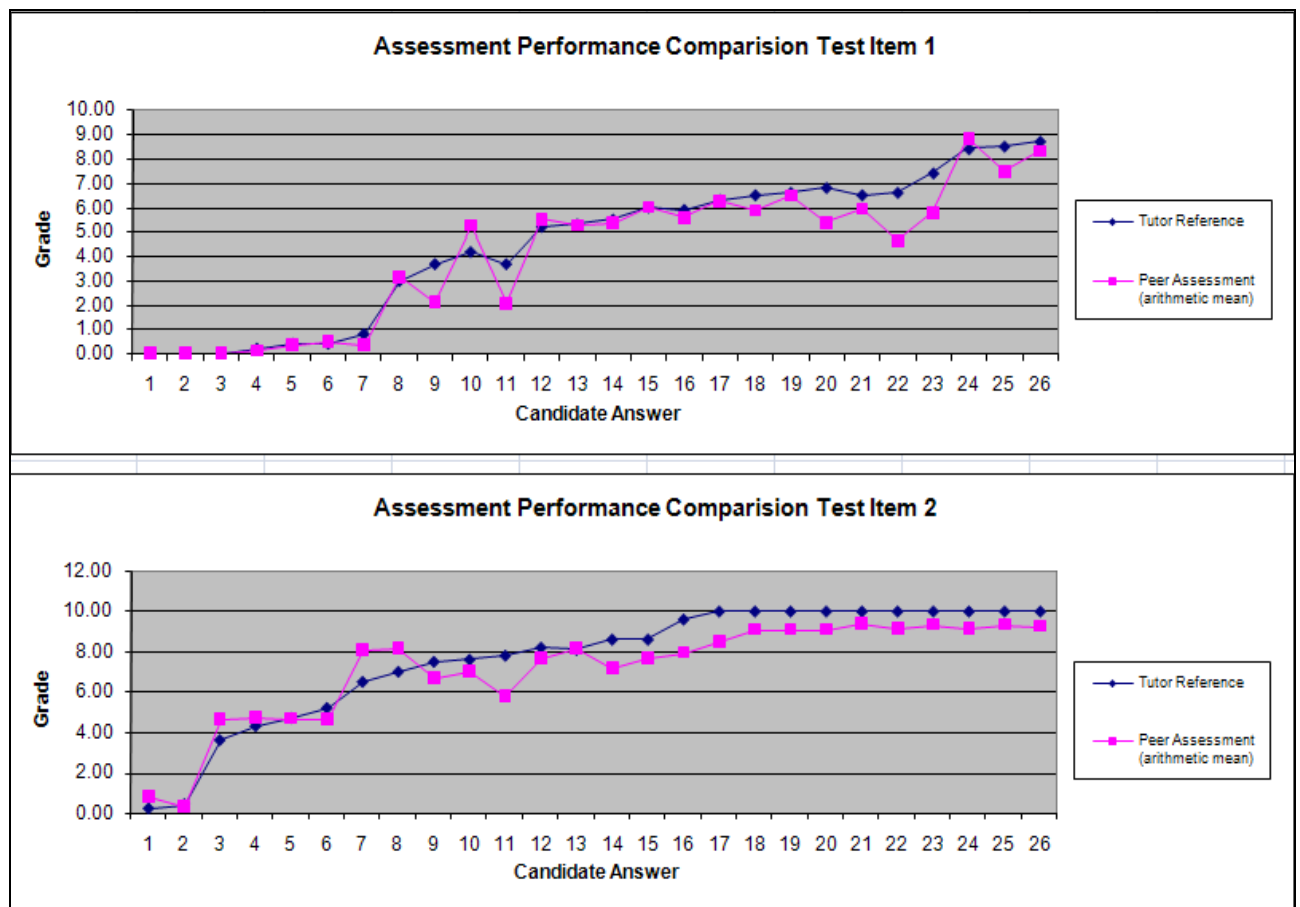
In order to investigate the results, the absolute error of the tutors' individual grading values is compared with the weighted average as in table 3. The absolute error for all of the test items is between 2.18 ($\sigma = 1.88$) as worst result and 1.12 ($\sigma = 1.14$) as best result. Similar to the cross-correlation findings, for test item 1 the absolute error varies between 2.30 ($\sigma = 1.84$) as worst result and 1.62 ($\sigma = 1.30$) 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 0.88 ($\sigma = 1.27$) and 0.45 ($\sigma = 0.62$). Test item 3 has not only lower cross-correlation but also higher absolute error values between 3.72 ($\sigma = 1.77$) and 1.07 ($\sigma = 0.80$). Because of students time limitations the candidate answers to be evaluated be the students have been split into two groups (11 for the first group and 14 for the second). The assessment performance between the two groups is quite similar and further investigations will be done in the future analysis.

Table 3: The absolute errors for tutor's assessment performance

	All Test Items		Test Item 1		Test Item 2		Test Item 3		Group 1		Group 2	
	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ
T1	1.12	1.14	1.83	1.42	0.45	0.62	1.07	0.80	1.42	1.19	1.24	1.23
T2	2.04	1.86	2.02	1.83	0.63	0.90	3.48	1.52	2.64	2.18	2.38	1.93
T3	2.18	1.88	1.93	1.36	0.88	1.27	3.72	1.77	1.97	1.75	2.25	1.79
T4	1.38	1.31	1.62	1.30	0.85	1.50	1.69	0.96	1.50	1.30	1.63	1.41
T5	1.65	1.69	2.30	1.84	0.69	1.57	1.95	1.23	1.89	1.66	1.85	1.75

4.2 Students Phase Results

In order to compare the student’s peer-assessment performance with the tutor’s reference marks, the arithmetic mean of peer’s individual results per candidate answer has been used and the absolute error of the student’s arithmetic mean and the tutor’s reference marks has been computed. For all the three test items (Q1, Q2, Q5) the arithmetic mean of absolute error is quite low with 0.98 ($\sigma = 0.95$). For the three test items individually, test item 1 has the lowest arithmetic mean of absolute error with 0.54 ($\sigma = 0.62$). Test item 2 has a value of 0.86 ($\sigma = 0.49$), where test item 3 has a higher value with 1.54 ($\sigma = 1.27$) which reflects the complexity of the assessment activity done by this item as an explanation one. The correlation between the arithmetic mean of the student’s candidate answer and the tutor’s reference grading for each candidate answer is quite high with 0.90 for all the three test items, 0.97 for test item 1, 0.96 for test item 2, and 0.86 for test item 3. Figure 1 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.



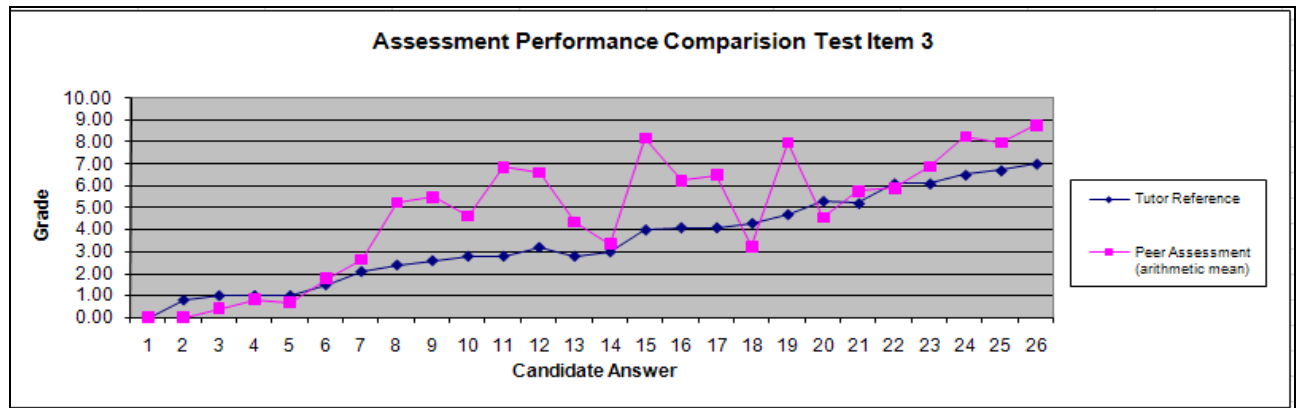


Figure 1: Students Peer-assessment Performance

5. Conclusion and future Work

In this paper, the reliability of the peer-assessment results was analyzed through an enhanced experiment procedure. The level of agreement between the student's peer evaluations and the tutor's reference grading values 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 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 assessment values and the students average marks for each candidate answer has been used to evaluate the performance of the students in the peer-assessment task.

For future work, the reliability of the assessment results will be further enhanced by analyzing the tagged sections from the candidate answers as correct, incorrect and irrelevant. The feedback activities will be also improved to provide both students and tutors with valuable information about the peer-assessment procedure. The web-based peer-assessment prototype will be improved according to the usability and functionality findings, as well as to the recommendations from the students and the tutors.

References:

- [1] Brown, G.; Bull, J.; Pendlebury, M. ,(1997): What is assessment?. In *Assessing Student Learning in Higher Education*. London: Routledge.
- [2] Bull, S.; Brna, P.; Critchley, S.; Davie, K.; Holzherr, C., (1999): The Missing Peer, Artificial Peers and the Enhancement of Human-Human Collaborative Student Modelling. In *proceedings of AIED 99*, 269-276.
- [3] (CPR)TM. Calibrated Peer Review, last retrieved Feb. 5th, 2009, from <http://cpr.molsci.ucla.edu/>
- [4] Davies, P. ,(2003): 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, Sheffield, UK.
- [5] Dochy, F. J.; & McDowell, L. ,(1997): Introduction. Assessment as a tool for learning. *Studies in Educational Evaluation*, 23 (4), 279-298.
- [6] Downing, T., Brown, I. ,(1997): Learning by cooperative publishing on the World-Wide Web. *Active Learning* 7, 14-16.
- [7] Eschenbach, E. A.; Mesmer, M. A. ,(1998): Web-based forms for design team peer evaluations. *American Society for Engineering Education 1998, Annual Conference and Exposition*, Session 2630.

- [8] Freeman, M.; McKenzie, J. (2002): 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.
- [9] Gehringer, E. F. (2000): Strategies and mechanisms for electronic peer review. In *Proceedings, Frontiers in Education Conference, Vol 1., F1B/2 - F1B/7.*
- [10] Gehringer, E. F. (2001): Electronic peer review and peer grading in computer-science courses, *Proc. of the Technical Symposium on Computer Science Education*, p. 139-143.
- [11] ISRDC. Document Classification. Online lecture as part of the course ISR, Graz University of Technology, last edited Nov 25th, 2007, last retrieved Sep. 29th, (2008) from <http://www.iicm.tugraz.at/isr/exp>
- [12] 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, last reviewed June 24th, 2009 from <http://www.ascilite.org.au/aset-archives/confs/edtech88/magin.html>
- [13] Ohland, M.W.; Loughry, M.L.; Carter, R.L.; Bullard, L.G.; Felder, R.M.; Finelli, C.J.; Layton, R.A.; Schmucker, D.G. (2009): The Comprehensive Assessment of Team Member Effectiveness (CATME): A New Peer Evaluation Instrument, *Proceedings of the 2006 ASEE Annual Conference, Chicago, Illinois, June 2006*. Information about CATME may be found at <http://www.catme.org>.
- [14] Orsmond, P. (2004): 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.
- [15] Rada, R.; Acquah, S.; Baker, B.; Ramsey, P. (1993): Collaborative Learning and the MUCH System. *Computers and Education*, 20(3), 225-233.
- [16] 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.
- [17] Sung, Y. T.; Chang, K. E.; Chiou, S. K.; Hou, H. T.. (2005): The design and application of a web-based self- and peer-assessment system. *Computer & Education*, 45 (2), 187-202.
- [18] Topping, K. (2003): Self and Peer Assessment in School and University: Reliability, Validity and Utility. In Mien Segers, Filip Dochy and Eduardo Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards*, Springer Netherlands, 55-87.
- [19] Topping, K. J.; Smith, E. F.; Swanson, I.; Elliot, A.(2000): Formative Peer Assessment of Academic Writing Between Postgraduate Students. *Assessment & Evaluation in Higher Education*, Vol. 25, No. 2, p. 150-169.
- [20] Trahasch, S..(2004): Towards a flexible peer assessment system. In *Proceeding, Information Technology Based Higher Education and Training (ITHET 2004)*, 516-520.
- [21] Tsai, C. C.; Lin, S. S.; Yuan, S.M. (2002): Developing science activities through a networked peer assessment system. *Computers & Education*, 38 (1-3), 241-252.
- [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.
- [23] WebPA, <http://webpaproject.lboro.ac.uk>. Last retrieved, 5th. February. (2009).
- [24] WPDC. Document classification. Wikipedia, last edited Sep. 8th, (2008), last retrieved Sep. 5th, (2009) from http://en.wikipedia.org/wiki/Document_classification

Author(s):

Mohammad AL-Smadi Ph.D. Candidate.

Institute for Information Systems and Computer Media, Graz University of Technology
Brückenkopfgasse 1, 8020 Graz, Austria
msmadi@iicm.edu

Christian Gütl Dipl.-Ing. Dr. techn. Univ.-Doz.

Institute for Information Systems and Computer Media, Graz University of Technology, Austria,
School of Information Systems, Curtin University of Technology, Perth, WA.
Infodelio Information Systems and GÜTL IT Research & Consulting, Austria.
Brückenkopfgasse 1, 8020 Graz, Austria.
cguetl@iicm.edu and cguetl@acm.org

Denis Helic Dipl.-Ing. Dr.techn. Univ.-Doz.

Institute for Information Systems and New Media, Graz University of Technology, Austria
dhelic@iicm.edu.