E-TESTER: a Computer-based Tool for Auto-generated Question and Answer Assessment

Christian Guetl
Institute for Information Systems and Computer Media (IICM) at Graz University of Technology,
Infodelio Information Systems, Internet Studio-Isser and GÜTL IT Research & Consulting, Austria
cguetl@iicm.edu and cguetl@acm.org

Heinz Dreher
School of Information Systems
Curtin University of Technology
Perth, Western Australia
h.dreher@curtin.edu.au

Robert Williams
School of Information Systems
Curtin University of Technology
Perth, Western Australia
bob.williams@cbs.curtin.edu.au

Abstract: Adaptive E-Learning systems are needed to efficiently support lifelong learning activities. To check goal attainment, learners need to take tests or assessment activities. Large efficiency gains may be made if the assessment and associated grading or evaluation process could be supported with adaptive systems sensitive to user parameters. Hand made tests are a time-consuming and tiresome task. If the tests are hand scored there is the inevitability of either interrupting the learning whilst the assessment outcome is pending, or the danger of proceeding with the attempt at learning possibly unsuitable content. As a contribution to solving the problem stated so far we have designed the E-TESTER. In this paper we describe the implementation solution and its integration into the AdeLE system, point out the potential of the E-TESTER by an application example, and discuss the experiences of the prototype implementation.

Keywords: adaptive E-Learning, knowledge acquisition, auto-generated questions, automatic answer assessment, AdeLE, E-TESTER, LMS, Learning Management Systems.

Background

E-Learning is a promising and widespread solution in a variety of application scenarios. According to the 2003 IDC study (IDC 2003), some 934 million USD was invested in E-Learning worldwide in that year.

Learners are seeking flexible study options which permit them to select content, experience learning activities consistent with a variety of individual characteristics, and take assessments best suited to their learning schedule as opposed to an educational organisation's schedule (i.e. exam period). This means an E-Learning system should support personalisation of learning goals, study content, learning pathways, navigation, and assessment such that an individualised learning package can be delivered to the student.

Existing LMS (Learning Management Systems) do support some of the above functions and to varying degrees, however there has been little support for assessment or evaluation. Typically, some Multiple-choice or True-false test is offered with little scope for adaptability – i.e. the tests are pre-determined and not dynamically created based on the material the student has studied in the associated session. Computer applications learning assessment software by Course Technology, for example SAM (2004), does provide a dynamic test environment and is an important development but is restricted to specific content. We wish to have a system which is independent of content, or rather, interoperates with any content.

Thus, an up-to-date and detailed user profile of the learner is required. Most common systems get their profile information from user-filled forms at the initial stage of system usage or by tracking page requests. For example learners' progress tracking implemented by the Hyperwave E-Learning Suite (HYPERWAVE 2004) tracks pages visited as a percentage of total pages and thus gives the learner an indication of quantity in the form of a progression bar. However, this may say precious little about knowledge acquisition.

To better represent learning progression we need a proper, fine-grained user profile for the adaptation process. Some interesting improvements have been made. For example the AdeLE system (García Barrios et al. 2004), which exploits real time eye-tracking and content tracking information, enables the system to get further profile information, such as learning style, topics of interest, and user preferences regarding scrolling vs pagination, hierarchical navigation vs exploration, etc. However, a significant problem in the context of this study is that eye-tracking and for example brain signal scanning and other sensory data will not give reliable hints about the learning acquisition. But especially this parameter would be important in the context of an adaptive E-Learning system in order to determine whether a lesson's objectives have been met, or if some repetition and learning reinforcement is required (Lennon & Maurer 2003; McLoughlin & Luca 2001).

Existing systems as well as contemporary research (see also "Computer-based Testing and Answer Evaluation" section below), use and trial some sort of manually generated tests to evaluate knowledge acquisition. For easy and automated evaluation, Multiple-choice or True-false tests are commonly used. Another interesting approach discussed by (Lennon & Maurer 2003) is to provide a crossword puzzle to the learner.

It is obvious that the creation of tests for knowledge acquisition measurement applicable for E-Learning is a time-consuming and troublesome task. When the adaptive and variable nature of the content is considered, the required effort will be multiplied. This led us propose and design E-TESTER, a computer-based system that automatically creates questions based on the dynamically selected E-Learning content provided to the user. Furthermore, E-TESTER also automatically assesses natural language keyboard-typed answers, can model the learner's concept acquisition, provide feedback regarding concept development, notably the degree to which the learner's acquired concepts match the learning objectives. To our knowledge no comparable system has yet been implemented.

The remainder of this paper is organized as follows: first an application example illustrates the potential of the E-TESTER, followed by a discussion of related work as well as a brief description of the E-Tester implementation solution and the integration into the AdeLE Framework. A discussion of the first experiences of the prototype implementation and future work conclude the paper.

Application Example

To point out the potential of the E-TESTER system in conjunction with an adaptive E-Learning system, the following application scenario is described. Irene has enrolled in the lecture "Copyright on the Internet (CORI)" using the adaptive E-Learning system encompassing E-TESTER in order to prepare herself for a new commercial project. The lecture CORI has four lessons from the lecture "Basics in copyright (BACOR)" as pre-requisites. Because of that, firstly the pre-requisite knowledge of these topics has to be examined. The E-TESTER system identifies the main concepts and creates a set of questions related to those four lessons. The E-Learning system delivers the questions and collects the natural language answers entered. E-TESTER assesses the answers and prepares an overview of the existing pre-requisite knowledge, and in the case of a deficit, the adaptive E-Learning system provides appropriate content to compensate for the weaknesses identified. Assuming the fact that Irene has a weaknesses in contract law, the necessary content will be provided. After acquiring the content and passing a further assessment, the CORI lecture will be started. Each lesson has a similar assessment process. In addition, this assessment information can be used to gain a fine-grained profile of acquired knowledge as well as weaknesses, which may be provided to the user in a graphic representation.

Computer-based Testing and Answer Evaluation

As already stated, testing and assessment are important in the application of E-Learning systems. The importance may be emphasized by existing standards, recommendations and frameworks, see for example (IMS Global Learning Consortium 2003) and (SCORM 2004).

Some E-Learning systems or E-Learning management systems provide tools for easy creation and management of assignments and assessment, see for example (Cristea & Tuduce 2004), (Pesin 2003), and (CORONET 2004). However, these tools are only supporting teachers and tutors in creation of hand made questions. Of course, automatic content abstraction and concept identification is an ongoing research topic since the early days of modern information retrieval, see for example (Sparck Jones & Willet 1997) and (Cardie 1997). But to our best knowledge, no tool has been invented to automatically create test questions based on content.

The support of assessment in E-Learning systems is difficult due to the variety of testing options in addition to the heretofore impossibility of text based scoring and assessment. Multiple-choice tests are often applied for system-supported, automatic assessment of knowledge acquisition (Kuechler & Simkin 2003). An advantage is perceived objectivity derived in part from the absolute consistency of scoring, however, this comes at the price of measurement accuracy. Another advantage is the obvious marking efficiency, once the answer template has been constructed. However, (Wood 1998) argues that free response questions, or *unseen text* in our terminology, may be superior in measuring achievement of educational objectives.

More complex is the assessment of natural language answers, exercises and essays. A simple but very helpful way is to provide a tool for managing student works as well as annotating and marking them, for example as implemented in the CORONET system (Dreher et al. 2004). However, an automatic assessment of natural language content is a big challenge. Some research work can be identified for automatic answer provision, for example (Lytinen et al. 2000), and for determining candidates of definitions by a given concept, see for example (Cimiano et al. 2004). In recent years there has been a growing interest in this problem of grading unseen textual input such as student essays.

First into the field seems to have been (Page 1966) with Project Essay Grade (PEG). Landauer et al. (1998) developed the Intelligent Essay Assessor, using Latent Semantic Aanalysis techniques. The E-*rater* (Burstein et al. 1998) scoring engine is now extensively used in the USA by the Educational Testing Service in its processing of the GMAT (Graduate Management Aptitude Test) exams. Another approach was taken by Larkey (1998) which used a text categorisation technique, text complexity features, and linear regression methods. Such developments and the interesting backgrounds to the research approaches can be referred to in (Shermis & Burstein 2003).

There is a productive research group at Curtin University in Perth, Western Australia, who are developing some new and innovative approaches to Automated Essay Grading (AEG) in addition to conducting trials of existing systems. Results of such trials have been reported in (Palmer et al. 2002) and (Williams & Dreher 2004). The E-TESTER was inspired by this work and enables auto-generated questions and the assessment of the answers as discussed in the following paragraph.

E-TESTER Implementation Solution and Integration in the AdeLE Framework

In order to achieve implementation flexibility E-TESTER is designed to be an encapsulated system which can interact with E-Learning and other systems. The system can be seen as a specialized service which is fed with content, generates questions and evaluates the answers against the content. Our application scenario is focussed on the measurement of knowledge acquisition which can be used for controlling the E-Learning session. For example the information may be applied to determine whether a student has passed an online lesson, or to identify a deficit in knowledge acquisition and thus the provision of additional content. Unlike *MarkIT* (Williams & Dreher 2004) E-TESTER is not designed to do automatic marking and rating, rather the system is focused on natural language text questions and answers, at this time in the English language. In the future, we would propose to accommodate other languages, assuming our approach is successful.

E-TESTER Implementation Solution

The idea is based on exploiting the E-Learning content as a model answer. That content, subdivided by predefined logical units (e.g. sections or lessons), is the source for the proprietary representation of the knowledge for further processing. The system uses an adapted Context Free Phrase Structure Grammar (CFPSG) parser to perform "chunking" of document sentences into Noun Phrase (NP) and Verb Clause (VC) structures. Proprietary structures, based on the rules of transformational grammar, are used to represent the semantics of the content. A thesaurus permits the construction of meta-level information. Many different expressions of the same document content can thus be represented by one semantic representation. It is the internal representation of these concepts which are further used by the E-TESTER system in two ways.

Firstly, the concepts represent candidates for questions. Based on these concepts the systems creates simple questions like "What is <*concept1*>?" or "Explain <*concept2*>." As a further work it is intended to extend the internal representation by semantic reasoning for creating more complex questions and for rating the importance of the concepts.

Secondly, the concept represents the model answer in an abstract way. By applying the same procedure as described above, the students' answers are processed to derive a semantic model. The assessment is done by comparing the concepts of the model answer and the concepts of the student answer.

E-TESTER Integration into the AdeLE Framework

AdeLE (Adaptive E-Learning with Eye-Tracking) is a four-year-funded research project by the Austrian government. The main objectives are to improve the knowledge acquisition applying adaptive E-Learning by advanced methods, see (García Barrios et al. 2004), and to provide an open and flexible framework for carrying out experiments and research approaches as well. Because of the last mentioned objective, the AdeLE implementation is based on a service-based architecture to guarantee an easy pluggable and changeable functionality, which is described elsewhere (Guetl et al. 2004). Following that concept, the E-TESTER system is also implemented as a service, which enables a widespread application of the system.

Because of its flexibility, the server side functionality of AdeLE is subdivided into 3 systems: (1) The E-Learning Management System (EMS) manages the curriculum and the learning assets as well as renders the personalized content and delivers it to the users' client. (2) The Adaptation System (AS) is responsible for estimating the personalized content, representation and navigation as well as controlling the EMS. (3) The Profiling System (PS) collects information of the users' behaviours, manages user models and provides the AS with proper and timely user information.

The integration of E-TESTER into an adaptive E-Learning environment is a great challenge, because of the personalized information delivery. For example, after finishing the basics of a lecture, students are allowed to select three advanced topics concerning their interests from a set of 12 topics. Such choice among elective topics must of course be reflected in the assessments E-TESTER delivers to the student, and is achieved via the module responsible for personalized assessment of knowledge acquisition. Within the AdeLE framework, E-TESTER is closely coupled to the AS.

According to the curriculum and the underlying course structure, the AS controls the assessment of knowledge acquisition as follows. AS requests the PS for information of personalized content delivered to the user. Based on that content, E-TESTER creates a set of possible questions, selects the requested number of questions randomly, and delivers the questions to the AS. AS forms a sequence of XML-form-based pages (each of which consist of a question and a text input form), and EMS delivers the pages to the user. The answers are sent back from the ELS to the AS and are assessed by E-TESTER against the content delivered to the user. Results are stored at PS for further usage and for controlling the workflow of further learning activities.

Prototype Implementation and First Experiences

The first running prototype implementation is not yet prepared to run as a service and is not fully integrated into the AdeLE system. The prototype is implemented in C++ and Java and permits identification of the main concepts from unseen content. This technique is used to create simple questions like "What is <concept>?". It also assesses the student answers against the course content.

First experiences show surprisingly good performance. Experiments have been conducted with a number of 1st year Information Systems student essays, and 2nd year Law student essays, both at university level, and also year 8 secondary school English essays. These essays were prepared by students using a word processor, and comprised some 300 to 500 words, or about one page of text. Expert human graders created the "Human" scores in the usual way by applying the model answer criteria to the essays presented to grading. The computer scoring was a rather simple process of compiling all student answers as text files and submitting it to the computer algorithm. Our technology takes less than 4 seconds to deal with the types of inputs described above. Providing the model answer, which is derived from the course content, to the computer is a slightly more involved task.

The graph in Figure 1 represents results for a sample of 20 essays (horizontal axis) in which the maximum possible assessment was 30 (vertical axis) and shows the comparison between expert human and computer assessments. The data is (arbitrarily) ordered by increasing computer score. Assignment 1 is assessed by the human at 2 and by the computer at 10 (leftmost data item). Assignment 10 at assessed at 21 by both human and computer, whereas assignment number 20 (rightmost data point) is assessed by the human at 27, and by the computer at 32 – yes, we omitted to inform the computer about the maximum mark on this run! As can be seen the computer tracks the human reasonably well, but further scoring algorithm refinement is indicated. For the purposes of E-TESTER, the evaluation challenge is simplified because we only need to check concept acquisition.

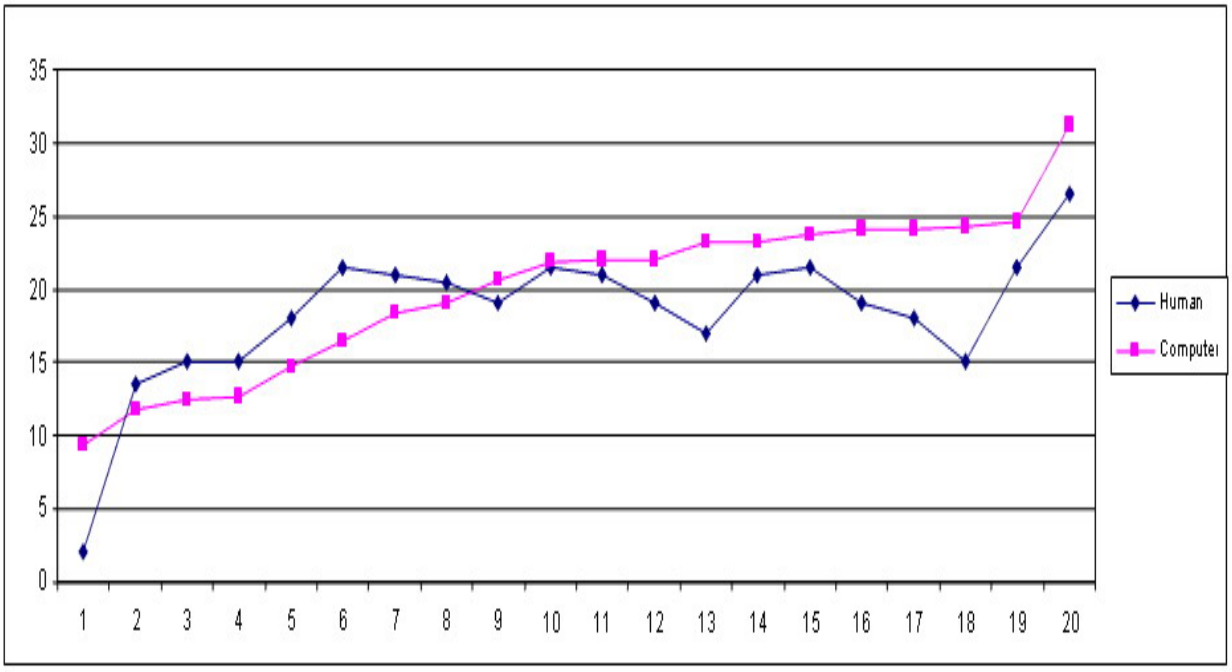

Figure 1 - Human vs computer-based scores in ascending order of Human scores

In Figure 2 below, we have presented another example of E-TESTER output. In this case we have a graph showing the 'concepts' associated with both the model answer and the student answer. Naturally, the better the correspondence between the concept representation in both, the better the score. If we focus on the tallest bar

(Concept_Number 31) we see that the student answer (dark bar) contains a concept_frequency of 6 (vertical axis) where the model answer called for no discussion on this topic or concept. We say the student has introduced irrelevancies into the answer; or perhaps this is what can be termed an error on the student's part. Concept_Number 26 has a better match between model and student answer, indicating the student has learned relevant material. There are three cases where the model answer concepts are not matched by a student contribution (3, 28, 30) – this we would call "ignorance" or a deficit in knowledge. Such visual feedback is rather informative to student and teacher alike. It is intended to further develop such visual feedback into a dynamic object which responds to inquiry for concept name (associated with Concept_Number), and the possibility of linking back to the sections of the student assignment which are good, and those needing improvement.

There are many other applications for this technology. Plagiarism is a growing problem in modern education systems, sophisticated forms of which can be potentially identified by our technology. To promote learning, hints could be provided for improvement, and most importantly, tailored study package content would be automatically compiled and offered for further study.

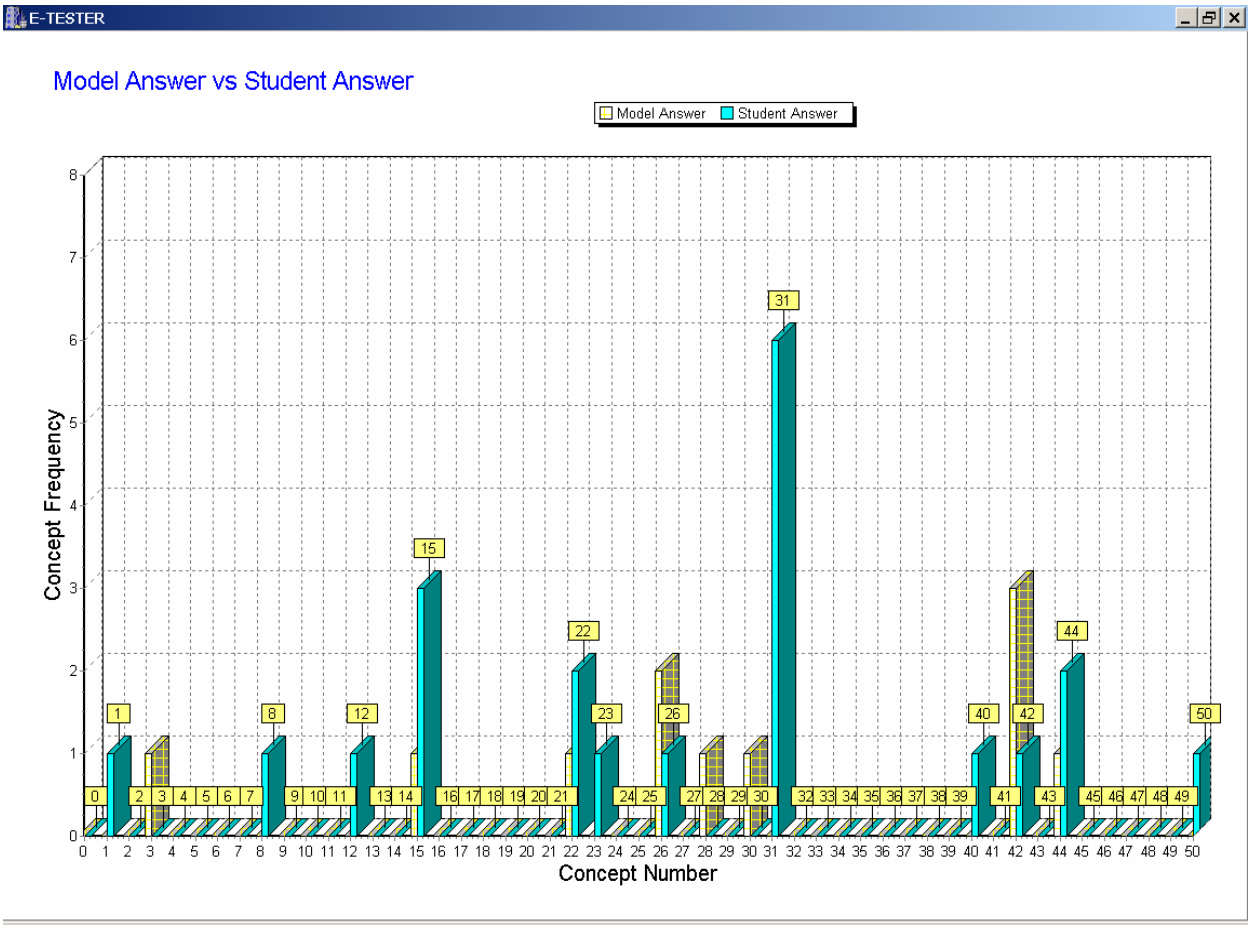

Figure 2 – Concept frequencies: student answer and course content

Future Work

The promising results of our first experiments led us to improve the prototype implementation by creating more advanced question types and by multilingual support. Also the E-TESTER module will be integrated closely into the AdeLE system by adding a web service interface. Furthermore, to enable the widespread use of E-TESTER, an open interface will be specified.

First experiences have also inspired us to do further research work. Interesting and also helpful will be a tool to carry out examinations applicable for E-Learning systems in distance learning. Another conceivable application for E-TESTER in Knowledge Management could be the computer-based generation of FAQ-type advice or content.

Furthermore, a learning paradigm which implements "learning by answering questions" could be investigated. Based on the course content the E-TESTER system creates some questions, and students have to work out answers (read the content, consult background information, find relevant content on the Internet, etc.). Finally, E-TESTER is able to assess the answers and provide feedback to the students. Another interesting research topic can be addressed by comparing the fine-grained user information gained by real-time eye-tracking information against the knowledge acquisition assessed by E-TESTER.

References

ADELE (2004) Website last visit 2005-03-13. http://adele.fh-joanneum.at/

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Enriching Automated Essay Scoring Using Discourse Marking, *Proceedings of the Workshop on Discourse Relations and Discourse Markers, Annual Meeting of the Association of Computational Linguistics*, August, Montreal, Canada.

Cimiano, P., Handschuh, S., & Staab, S. (2004). Towards the Self Annotating Web, *Proceedings of 13th World Wide Web Conference*, 17-22 May, New York, USA. http://www2004.org/proceedings/docs/1p462.pdf

Cardie, C. (1997). Empirical Methods in Information Extraction; *AI Magazine*, Vol. 18, No. 4, 1997, pp.65-80. http://www.cs.cornell.edu/home/cardie/papers/ai-mag.pdf

Cristea, P.D., & Tuduce, R. (2004). Test Authoring For Intelligent E-Learning Environments, *First International Workshop on Authoring of Adaptive and Adaptable Educational Hypermedia*, paper no. 416-800. Technical University Eindhoven. http://wwwis.win.tue.nl/~acristea/WBE/416-805_WBE-PCristea_RTuduce_6pg.pdf

CORONET (2004). Website last visit 2005-03-13. http://coronet.iicm.edu/

Dreher, H., Scerbakov, N., & Helic, D. (2004). Thematic Driven Learning. *Proceedings of E-Learn 2004 Conference*, Washington DC, USA, November 1-5, 2004.

García Barrios, V., Gütl, C., Preis, A., Andrews, K., Pivec, M., Mödritscher, F., & Trummer, C. (2004). AdeLE: A Framework for Adaptive E-Learning through Eye Tracking, *Proceedings of IKNOW 2004*, Graz, Austria, 2004, pp.609-616.

Gütl, C., García Barrios, V., & Mödritscher, F. (2004). Adaptation in E-Learning Environments through the Service-Based Framework and its Application for AdeLE. *Proceedings of E-Learn 2004 Conference*, Washington DC, USA, November 1-5, 2004.

Hyperwave (2004). Hyperwave eLearning Suite, Website last visit 2005-03-13. http://www.hyperwave.com/e/products/els.html

IDC (2003). *U.S. Corporate and Government eLearning Forecast. 2002-2007*. Website last visit 2005-03-13. http://www.idc.com/getdoc.jhtml?containerId=30119

IMS Global Learning Consortium (2003). *IMS Question & Test Interoperability Specification*. Website last visit 2005-03-13. http://www.imsglobal.org/question/index.cfm

Kuechler W. L., & Simkin M. G. (2003). How Well Do Multiple Choice Tests Evaluate Student Understanding in Computer Programming Classes? *Journal of Information Systems Education*, Winter 2003, Vol. 14(4).

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis, *Discourse Processes*, Vol. 25, pp.259-284.

Larkey, L. S. (1998). Automatic Essay Grading Using Text Categorization Techniques, *Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp.90-95.

Lennon, J., & Maurer, H. (2003). Why it is Difficult to Introduce e-Learning into Schools and Some New Solutions; Journal of Universal Computer Science, Vol. 9, No. 10.

Lytinen, S. L., Tomuro, N., & Repede, T. (2000). The use of WordNet Sense Tagging in FAQFinder. *Proceedings of the AAAI-2000 workshop on AI and Web Search*, Austin TX, July 2000.

McLoughlin, C., & Luca, J. (2001). Quality in Online Delivery: What does it mean for assessment in E-Learning environments? *Meeting at the Crossroads: Proceedings of the Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCLITE 2001)*, Melbourne, Australia, 2001.

Page, E. B. (1966). The Imminence of Grading Essays by Computer, *Phi Delta Kappan*, January, pp.238-243.

Palmer, J., Williams, R., & Dreher, H. (2002): Automated Essay Grading System Applied to a First Year University Subject – how can we do it better? *Proceedings of Informing Science 2002 Conference*, Cork, Ireland, June 19-21. http://proceedings.informingscience.org/IS2002Proceedings/papers/Palme026Autom.pdf

Pesin, L. (2003). Knowledge Testing and Evaluation in the Integrated Web-Based Authoring and Learning Environment. *Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies (ICALT'03)*. http://ieeexplore.ieee.org/iel5/8621/27318/01215077.pdf

Shermis, M., & Burstein, J. (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, New Jersey, USA.

SCORM (2004). *Advanced Distributed Learning SCORM Overview*. Website last visit 2005-03-13. http://www.adlnet.org/index.cfm?fuseaction=scormabt

Sparck Jones, K., & Willet, P. (1997). *Readings in Information Retrieval*. Morgan Kaufmann Series in Multimedia Information and Systems.

SAM (2004). Thomson Course Technology. Website last visit 2005-03-13. http://www.course.com/catalog/product.cfm?isbn=0-619-17345-9&CFID=4626963&CFTOKEN=82851351

Williams, R., & Dreher, H. (2004). Automatically Grading Essays with Markit© *Proceedings of Informing Science 2004 Conference*, Rockhampton, Queensland, Australia, June 25-28, 2004. Website last visit 2005-03-13. http://proceedings.informingscience.org/InSITE2004/092willi.pdf

Wood, W. C. (1998). Linked Multiple-Choice Questions: The Tradeoff Between Measurement Accuracy and Grading Time; *Journal of Education for Business*, Nov/Dec 1998, Vol 74, No. 2.

Acknowledgment

The AdeLE project is partially funded by the Austrian ministries BMVIT and BMBWK, through the FHplus impulse programme. The support of the following institutions and individuals is gratefully acknowledged: Department of Information Design, Graz University of Applied Sciences (FH JOANNEUM); Institute for Information Systems and Computer Media (IICM), Faculty of Computer Science at Graz University of Technology; especially Hermann Maurer and Karl Stocker as well as all other members of the AdeLE team, in particular Maja Pivec, Victor García Barrios, Felix Moedritscher, Christian Trummer, Juergen Pripfl and Martin Umgehe.

The Curtin University research effort has been funded in part by grant monies from the Teaching and Learning Network and Curtin Business School.