Online Educa Berlin, 04.12.2008

Dynamic Test Generation in e-Assessment: A way to handle mass assessment while keeping quality?

Dr. Jens Bücking, Zentrum für Multimedia in der Lehre (ZMML) – Universität Bremen, Germany

The rapidly increasing number of exams in higher education, driven by the bologna-process, is a challenge not only for students but also for teachers and awarding bodies at universities. The workload of teachers can be substantially reduced by automatic or semi-automatic scoring methods in e-assessments[1] (Fig.1). However, to keep quality of testing, more time is needed for the development and quality management of digital questions, especially for innovative items integrating multimedia elements like sound, video and interactive simulations.

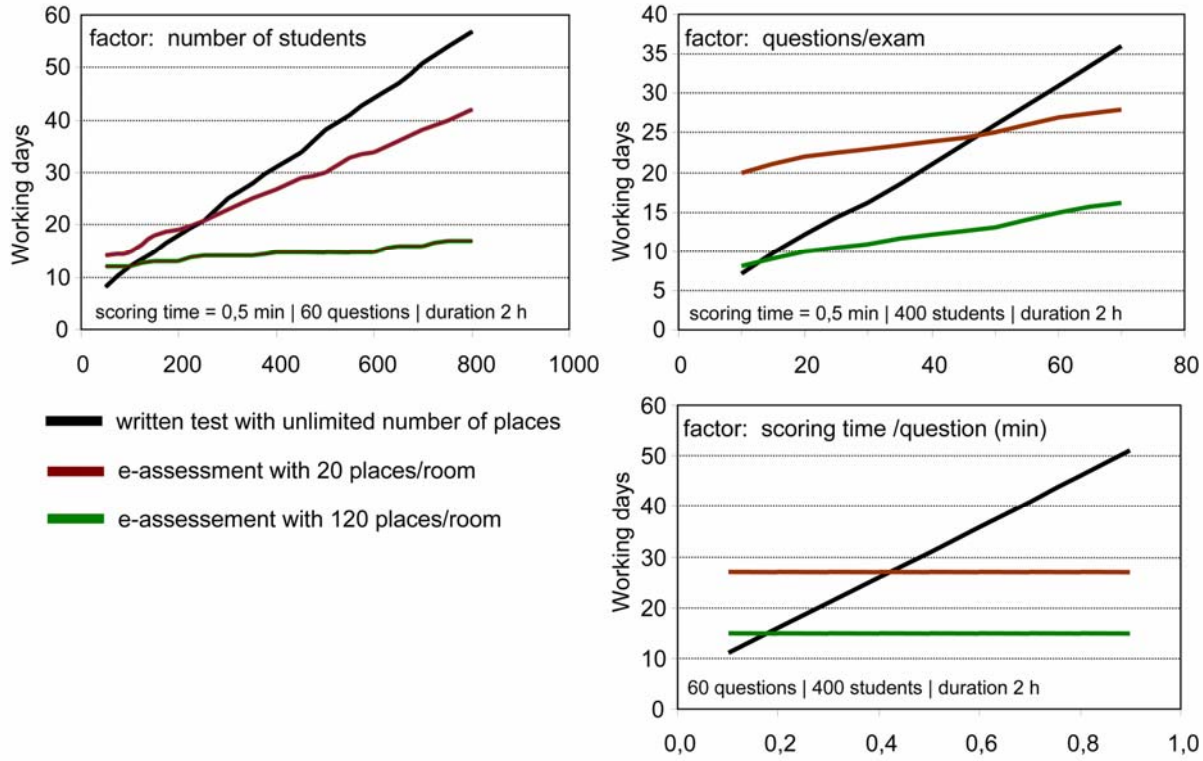

Fig. 1: Work load in comparison of on-paper and high stake e-assessments

Fig.1 is based on a mathematical model summing up the workload for production and quality management of questions, preparation of the exam, supervision of exams and scoring (on-paper asessments only).

The “Center for Multimedia in Higher Education” (ZMML) at the University of Bremen is using the LPLUS TestStudio (LPLUS GmbH, Bremen - Germany) for high stake assessments. In the context of the e-Learning service „e-assessment“ since February 2005 more than 16.000 computer based exams took place. Actually this examination method is used in 8 of 12 faculties with 3-4 thousand subscribed participants each semester.

The introduction of e-assessments often is accompanied by prejudices, i.e (i) that automatically scored tests are composed only of multiple choice questions and (ii) that multiple choice questions can only be used to test the repetition of knowledge. Both statements are wrong. E-assessments are capable to integrate all features offered in web based trainings. Video analysis and sound integration, e.g. to test listening comprehension, are only two of many examples. The level of competence tested by multiple choice questions can also be application, analysis or transfer of knowledge. The main pedagogical risk in implementing e-assessments is, that the time teachers safe by automatic scoring will not, at least partly, be reinvested in developing high quality questions.

[1] The terminology of the abstract follows the JISC and QCA e-assessment glossary (see http://www.jisc.ac.uk/uploaded_documents/eAssess-Glossary-Extended-v1-01.pdf)

The implementation of the e-Learning-service for e-assessments at the University of Bremen, with some thousand electronic exams per semester, had to cope with an insufficient and distributed IT-infrastructure with heterogeneous hardware and software facilities. The work load for technical preparation, ensuring security and safety of data, ensuring high availability of the test system and supervising the exams in such conditions reduced the rationalisation effect of e-assessments and inhibited the campus-wide role out of the service. The number of interested teachers and faculties, however, is still increasing. As a solution to this problem, in December 2007 the University of Bremen opened its test centre, a highly specialised computer pool with 120 work places in a closed, high performance network environment.

Beside the need to setup an adequate IT environment, concepts for an optimal utilization of the existing capacities must be developed. If the number of participants exceeds the number of workplaces, more than one session is needed. Using the same static set of questions, a maximum of two sessions can be organised without the risk that questions are passed to the next participants. The exams organised by the ZMML, i.e. those before the opening of the test centre, had up 900 participants, with a mean number of 30 workplaces/room. With a mean of 210 registered students 9 sessions/exam were necessary. To avoid test exposure, huge item banks had to be developed, containing appr. 4 times more questions than beeing choosen randomly for each participant.
Fig. 2 shows the mean scores participants reached in 6 subsequent session of the main exam (August 2006) and in 6 sessions of the first repition (participants who failed or missed the main exam). In all sessions the same item bank was used with a random choice of 20 of 140 questions.

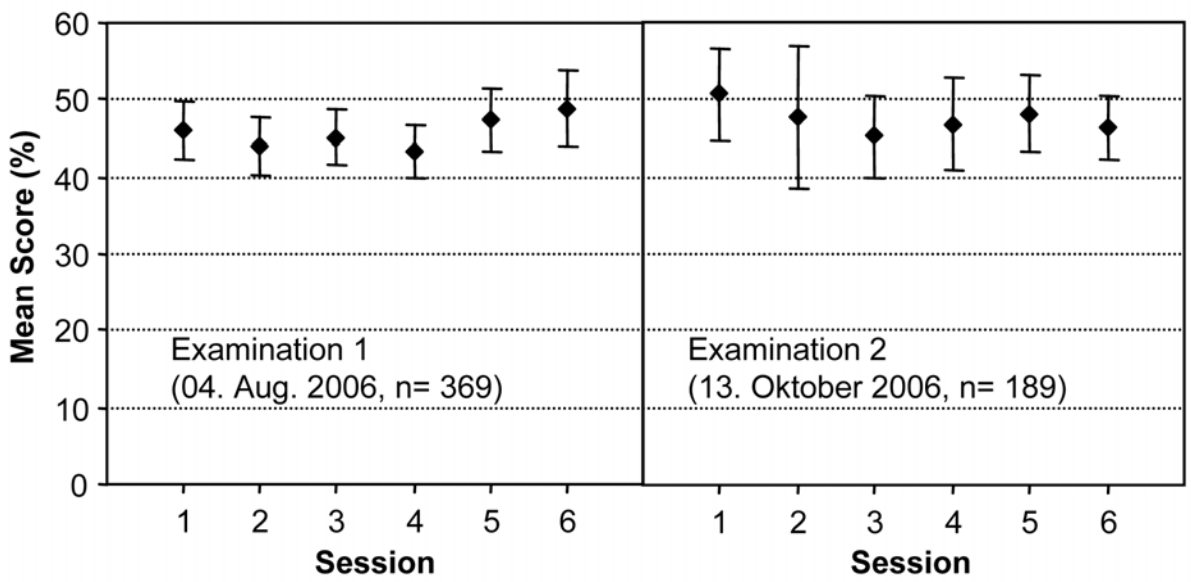

Fig. 2: Comparison of mean raw scores achieved by participants in an exam on business sciences. The exam was organised in 6 subsequent time slots. The error bars indicate the confidence interval with alpha = 0.05.

The low difference of the mean values indicate, that the influence of passed questions on the results is below statistical relevance. This does not exclude, that in some cases the result of the exam of individual participants have been influenced. The author will discuss this problem in comparison to on-paper assessments.

This dynamic test generation takes for granted, that pooled questions are equal in degree of difficulty and field of competence. If using the question pool for the first time, this can be decided only by the author of the question. Later on, this dicision can be supported by statistical methods analysing the examination results. In few cases this analysis led to the exclusion of single questions from the overall rating and to a modification of the item bank before beeing used for the next exam.

The development and quality management of huge item banks is new to many teachers and must be fostered by sufficient support structures and features of the test software. The presentation will show the development and quality cycles established at the University of Bremen using the web based software LPLUS TestStudio®. This process involves web based evaluations of the item bank, feedback of students given during the exam and the statistical validation after each usage of the item bank.