An Automated Individual Feedback and Marking System: An Empirical Study

Trevor Barker
University of Hertfordshire, Hatfield, UK
t.1.barker@herts.ac.uk

Abstract: The recent National Students Survey showed that feedback to students was an ongoing problem in Higher Education. This paper reports on the extension of our past research into the provision of automated feedback for objective testing. In the research presented here, the system has been further developed for marking practical and essay questions and providing automated feedback. Recent research at the University of Hertfordshire was able to show that learners and tutors accept and value our automated feedback approach based on objective tests and Computer Adaptive Testing. The research reported in this paper is an important extension to this work. The automated feedback system developed for objective testing has been extended to include practical testing and essay type questions. The automated feedback system, which can be used within any subject area, is based on a simple marking scheme created by the subject tutor as a text file according to a simple template. Marks for each option and a set of feedback statements are held within a database on a computer. As marks are awarded for each question by the teacher an individual feedback file is created automatically for each learner. Teachers may also add and modify comments to each learner and save additional feedback to the database for later use. Each individual feedback file was emailed automatically to learners. The development of the system is explained in the paper and testing and evaluation with 350 first year (1 final practical test), 120 second year (1 written and 1 practical tests) and 100 final year (1 final practical test) undergraduate Computer Science students is reported. It was found that the time to mark practical and essay type tests was reduced by more than 30% in all cases compared to previous years. More importantly it was possible to provide good quality individual feedback to learners rapidly. Feedback was delivered to all within three weeks of the test submission date. In end of module tests it was very beneficial indeed as it had proven difficult to provide feedback in the past after modules had ended. Examples of the feedback provided are presented in the paper and the development of the system using a user-centred approach based on student and staff evaluation is explained. The comments of staff teaching on these modules and a sample of students who took part in this series of evaluations of the system are presented. The results of these evaluations were very positive and are reported in the paper, showing the changes that were made to the system at each iteration of the development cycle. The provision of fast effective feedback is vital and this system was found to be an important addition to the tools available.

Keywords: assessment, feedback, automated systems, development, evaluation

1. Introduction

High staff/student ratios often mean that tutors often have great difficulty in providing students with high quality feedback on assessment performance that is timely and meaningful. This is despite the many advances in computer aided assessment and technology. Chickering and Gamson (1987) list ‘prompt feedback’ amongst their seven recommendations for good practice in teaching. Promptness is an important factor but other factors are equally so. For example feedback must be constructive, appropriate, useful, accurate, individual, delivered in context, detailed and should also facilitate feed-forward. Freeman & Lewis (1998) amongst others have reported on the importance of feedback as a motivator for student learning. Given the increasing pressures on teachers’ time, these goals are becoming increasingly difficult to achieve. Thus, there is an increasing demand for the development of software applications that would enable the provision of timely, individual and meaningful feedback to those learners. In previous work the author and colleagues have reported on the use of automated feedback systems related to Computer Adaptive Testing (CAT) (Barker, 2009; Lilley et al, 2004). In this research it was found that the systems developed had many of the benefits listed above. Constructive detailed feedback was delivered quickly, accurately, in context and it was possible to facilitate appropriate feed-forward for individual learners. In addition feedback on the cognitive level at which learners were effectively working was provided, related to Bloom’s taxonomy (Bloom, 1956; Anderson & Krathwohl, 2001). CAT systems were tested and evaluated by staff and students and shown to be effective and highly valued (Lilley et al., 2004). For feedback to be effective, it is argued, it should be individual for each learner and timely. The use of our feedback system based on objective computer-based adaptive testing was shown to be effective, but limited as it could only be applied to CATs. The CAT applications that were used in our feedback test systems have been reported by Lilley and colleagues (Lilley & Barker, 2002; 2003; 2004; Lilley et al., 2004; 2005). However it is important that fast and effective feedback be provided for a wider range of tests, especially practical tests which occupy a significant amount of the assessments in the domain of Computer Science and
also essay and written text questions. These are not catered for at all in CATs. The work described in this paper therefore relates to the development, testing and use of an automated marking and feedback system for essay and practical assessment.

1.1 Feedback provision

The use of technology in order to support learning has been shown to be highly regarded and expected by learners (Parkin & Thorpe, 2009). There is evidence from previous research in the literature that many students expect to receive their grades and feedback online using the affordances that technology brings to learning (Bloxham & Boyd, 2007). Hepplestone and Mather (2007) provide supporting evidence for the importance of providing online feedback to learners via their ‘feedback wizard’. Students value the flexibility, privacy and convenience of receiving feedback in this way. Price & O’Donovan (2008) suggest that providing feedback in this way engages students and allows them to respond to feedback at a time when they are emotionally prepared to do so. The timeliness of feedback has also been stressed in the literature, for example Mutch (2003) emphasizes the importance of providing feedback at a time and in a context when it is still meaningful to learners. This is particularly important in the provision of feed-forward in order to guide future learning and preparation for future assignments. Winter & Dye (2004) have shown that students are less likely to collect their feedback and grades unless they are provided within a reasonable time from the assignment. The format of feedback has also been shown to be important in studies. Feedback that is typed, and in a clear and legible format has been shown to be more acceptable to learners than other forms, for example hand written feedback (Bridge & Appleyard, 2005; Denton et al, 2008). Feedback system that deliver well formatted text via an online system were likely to be more highly valued by learners. Automated systems would also be expected to be more efficient and reduce the time taken by tutors to mark work (Jones & Behrens, 2003). Based on the evidence from the literature summarized here, it was decided to develop an automated feedback system that would deliver typed and formatted feedback to learners in a fast and secure fashion using electronic mail.

1.2 Methodology

The rapid development of computer software systems is readily facilitated using a user-centred prototyping approach (Sommerville, 2010). Prototyping has several benefits according to Sommerville. These include the clarification of requirements, providing a focus and direction for designing the system, and importantly the involvement of the stakeholders in the project and end users of the system (Sommerville, 2010). It also allows the system developer insight into the accuracy of initial project objectives and whether they can be successfully met. In complex domains such as those found in education, a user-centred prototyping approach is essential in order to evaluate the effectiveness of the system in a full context (Barker & Barker, 2002). In the development of the automated feedback system described here, expert users (tutors and assignment moderators) and small groups of end users (student recipients of the feedback) were employed in order to guide the prototype iterations. Nielsen and Mack (1994) have shown that small groups of expert users are able to facilitate rapid and efficient evaluation of computer systems in this way in order to support rapid prototyping.

2. Requirements of the system

In complex domains such as teaching and learning, the evaluation of implemented systems by stakeholders at all stage of the development process is absolutely vital as explained by Barker and Barker (2002). Teachers and learners as well as other stake-holders were expected to have a significant input into the nature of the system developed. The first stage of this was to develop a set of requirements for the system that would enable implementation, testing and subsequent improvements to the system. The initial requirements of the proposed system therefore, were arrived at as follows. Based on a survey of the literature summarized above and the evaluation of previous automated feedback system (Lilley et al., 2004) which included significant input from staff, students and academic managers a list of desirable functional requirements for the proposed system was produced. After discussion with colleagues and modification a basic set of functions was produced for the design of a first stage prototype. These are presented in table one. The ten functions shown in table one, were considered to be the minimum set necessary in order to develop the first stage prototype. This prototype would be developed and the implementation tested and evaluated in a real context. It was intended then that the results of this evaluation would enable the production of an improved set of requirements based on this experience.
Table 1: List of agreed functional requirements for the feedback system

<table>
<thead>
<tr>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>The system should be a computer-based marking system.</td>
</tr>
<tr>
<td>Simple to install and useful for a range of assessments and assessment types.</td>
</tr>
<tr>
<td>It should be able to mark both practical and essay type questions.</td>
</tr>
<tr>
<td>It should provide fast feedback.</td>
</tr>
<tr>
<td>The list of students and email addresses to be read in from university admin system in order to minimize work for the teacher.</td>
</tr>
<tr>
<td>Teachers would be able to enter five levels of feedback for each question.</td>
</tr>
<tr>
<td>General feedback would be allocated for each question based on the mark awarded.</td>
</tr>
<tr>
<td>The system would collate marks and produce feedback records for teachers.</td>
</tr>
<tr>
<td>Individual feedback and marks for each learner to be saved to a database file.</td>
</tr>
<tr>
<td>Feedback and marks to be distributed via electronic mail after checking.</td>
</tr>
</tbody>
</table>

3. Development cycles

The first prototype was developed using a standard Microsoft event driven programming language. This was decided upon mostly for speed and for ease of installation and testing. The system consisted of three main parts. A feedback file that contained the general feedback for each question, a student file that contained the list of students and their details, provided by the university admin system and a graphical user interface that read in the feedback and student files in order to allocate marks and feedback. The output from the system was a file which contained marks and feedback suitable for distribution via electronic mail. This was achieved by using a simple mail merge application within a Microsoft word processing application that read the file and applied it to a mail merge template developed for this purpose. Figure one shows the first prototype developed in this study.

![First version of the automated feedback and marking system](image)
The prototype shown in figure one above was used to mark a summative practical test for a set of approximately 350 first year computer science students. This test was taken under supervised conditions in a computer laboratory. Practical work was uploaded to the University’s managed learning environment (MLE) for marking. The test consisted of 16 questions. The text box below the buttons presents performance indicators for the mark. The buttons are used to allocate feedback for each question and the actual mark awarded for each section was entered by the marker after the appropriate button had been selected. An example of the feedback provided for each mark range is shown in table 2 below.

3.1 Format of the feedback

The feedback file created for use in the prototype was developed based upon a marking scheme for the assignment. For each question in the assignment five general feedback statements were written for excellent, good, fair, poor and absent or below acceptable standards. After discussion between markers these were manually written into file which suitable for reading by the prototype. Table two below shows an example of the feedback provided for one question.

Table 2: Example of feedback range provided for one question

<table>
<thead>
<tr>
<th>Part 4: 10% Sequence of still photographs</th>
</tr>
</thead>
<tbody>
<tr>
<td>10: Excellent sequence of stills with excellent subject matter and high quality images</td>
</tr>
<tr>
<td>7.5: Good sequence of stills with good subject matter and good quality images</td>
</tr>
<tr>
<td>5: Fair sequence of stills with some issues with either subject matter or the quality of images</td>
</tr>
<tr>
<td>2.5: Poor sequence of stills with considerable issues with either subject matter or the quality of images</td>
</tr>
<tr>
<td>0: Sequence of stills was not present or below acceptable standard</td>
</tr>
</tbody>
</table>

The number in each section is a guide to the marks relating to each of the feedback comments.

3.2 Use of the prototype

The first stage prototype shown in figure one above was used to mark approximately 350 practical assignments over a one week period. After the marking was completed marks were transferred to a spreadsheet in order to check that no errors had been made in marking, markers were consistent and that the mean and other statistical measures for the test were similar to other tests on the module. A sample of marked work and feedback was then passed to an external marker to be moderated and his comments were received for later analysis.

Once the course team was satisfied that the test had been marked fairly and accurately, marks and feedback were released via electronic mail to individual learners. In previous years it had proven difficult to achieve this timescale with smaller groups of approximately 250 students within a six week period. On this occasion we were able to release the marks three weeks after the end of the assignment for a group of 350 learners. Markers reported that the marking itself was faster and more efficient, taking approximately 30% less time to mark the work than previously. The greatest saving on time was related to writing and distributing feedback. On some occasions in the past it had not been possible to deliver feedback until after the end on the course itself and on one occasion feedback was not delivered at all since students were on their summer vacation by the time feedback was ready for distribution.

4. Evaluation of the first prototype

Approximately one week after the marks and feedback had been distributed, markers met to discuss and reflect upon the exercise. Comment from the external marker were also distributed and considered. Fifteen students were selected quasi-randomly to answer a short questionnaire. Selection was based on their scores obtained in the test. It was important that students with a range of scores in the test had an opportunity to comment on the feedback provided by the system, so five students were selected in each on the performance ranges, under 50, between 50 and 75 and above 75 marks. Table three presents a summary of their responses to the questionnaire.
Table 3: Learners’ responses to questionnaire and score achieved in the test

<table>
<thead>
<tr>
<th>Statement</th>
<th>Likert Scale (1 to 5)</th>
<th>Student responses (n=15)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>score &gt; 75</td>
<td>score 50-75</td>
</tr>
<tr>
<td>Feedback was useful to me</td>
<td>4.0</td>
<td>3.8</td>
</tr>
<tr>
<td>Feedback was fast</td>
<td>4.8</td>
<td>4.6</td>
</tr>
<tr>
<td>Feedback was delivered conveniently</td>
<td>4.8</td>
<td>4.2</td>
</tr>
<tr>
<td>Feedback was fair</td>
<td>4.4</td>
<td>3.8</td>
</tr>
<tr>
<td>The amount of feedback was good</td>
<td>4.2</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Several issues however were raised by markers and the external moderator relating to the feedback and functions available in first stage prototype. Perhaps the most important related to the flexibility of the feedback provided. Although markers considered it extremely useful it was considered to be very inflexible. Feedback comments of the type shown in table two above were considered to be too general and inflexible by the markers. For example in table two the feedback statement “Excellent sequence of stills with excellent subject matter and high quality images” relates to image quality and subject matter. It would be an improvement to separate this into two sections and provide a mark and appropriate feedback for each

Markers also wanted to add their own feedback comments on the assignment related to aspects of the performance overall. It was also considered useful to include some feedback related to the completeness of the work handed in. These two features were added to the list of functional requirements presented in table one to be used in the development of the next stage prototype.

In order to make the feedback more relevant, marking schemes were re-written in such a way that feedback could be related more specifically to performance. This was achieved by breaking each of the questions into smaller parts and writing the marking scheme and feedback comments to reflect this more closely. In this way a larger amount of more directed feedback could be written relating to each section of a question.

Another suggested improvement was the replacement of the rather inefficient way in which marks and feedback were allocated using the buttons shown in figure one. It was suggested that marks be awarded each section of a question and that feedback would be presented based on the mark awarded. In the previous system feedback was awarded by selecting the appropriate button and the mark entered later. This modification would do away with the buttons altogether and make the use of the system more efficient. It was also decided to produce additional introduction and summary screens to show in the first place the submission requirements for each assignment and also a final screen summarizing the marks and feedback for each learner.
5. Development of the second prototype

The modifications were made to the system as outlined in section 4 above and a second stage prototype of the system was developed and tested prior to use with students on an assignment. The modified version of the system allowed markers to comment on the completeness of the hand-in for the assignment as shown in figure two. In this version, the hand-in information is presented to the marker who may then make additional comments on the completeness or nature of the hand-in.

![Figure 2: Modified version of the system, showing hand-in information and tutor-entered feedback](image)

Each of the questions and question parts were presented by the system and the marker was simply required to enter a mark. Appropriate automated feedback was determined by the system based on the mark awarded in each section of a question, reading it from the database file for the assignment. After all the question sections had been marked, the system presented a final summary screen so that the marker could check that the marks had been awarded accurately. If this were not the case, the marker could cancel the entry and mark the student again. An example of the summary screen is shown in figure three.

6. Evaluation of the second prototype

The second prototype was used to mark three assignments, a second year practical summative assessment for 125 students and two final year summative assessments for approximately 100 students. These assignments were complex in that they contained both practical and theoretical elements and were completed over a six week period. In the past it had been extremely difficult to achieve consistent marking in this type of assignment. They were slow to mark and feedback delivered in terms of quality and quantity was fairly inconsistent between markers.
After the assignments had been marked, moderated and subject to quality measures, marks and feedback were released as before. In this case, a sample of 15 students only from the second year module was selected using the same sampling method as before in order to test student attitude to the feedback using the same questionnaire as before. The results of this are shown in table four below.

Table 4: Second year learners’ responses to questionnaire and scores achieved in the test

<table>
<thead>
<tr>
<th>Statement</th>
<th>score &gt; 75</th>
<th>score 50-75</th>
<th>score &lt; 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback was useful to me</td>
<td>4.8</td>
<td>4.2</td>
<td>4.0</td>
</tr>
<tr>
<td>Feedback was fast</td>
<td>4.8</td>
<td>5</td>
<td>4.6</td>
</tr>
<tr>
<td>Feedback was delivered conveniently</td>
<td>5</td>
<td>4.8</td>
<td>4.8</td>
</tr>
<tr>
<td>Feedback was fair</td>
<td>4.6</td>
<td>4.4</td>
<td>4.4</td>
</tr>
<tr>
<td>The amount of feedback was good</td>
<td>4.8</td>
<td>4.6</td>
<td>4.4</td>
</tr>
</tbody>
</table>
As before, the attitude of learners to the feedback was positive. The small sample size (n=15) was chosen in order to facilitate rapid development of the prototype rather than to identify significant differences between the means displayed in table 4. It was therefore not possible to perform a statistical analysis on these results. However cautious comparison of the results displayed in table 4 suggests that learners were slightly more satisfied with modified prototype than with the earlier version. A larger study to investigate the significance of any differences in attitude related to performance will be undertaken when the final version of the prototype becomes available.

Tutors were also invited to discuss their experiences of using the second prototype system and these were taken along with comments from the external marker. It was generally agreed that the system was improved by the modifications in terms of efficiency of marking and the quality of feedback presented to learners. As before the system performed faultlessly and there were no problems with installation or the automated distribution of feedback via electronic mail. One issue that surface is that although learners were in general satisfied with the feedback awarded, there was a greater number of students prepared to challenge their mark. It is suggested that this was due to the more detailed and specific marking scheme provided along with more detailed feedback relating specifically to each question part. Errors in marking made by markers were more readily identified by students and these were naturally more likely to be questioned. This issue was considered to be a positive feature by markers of the system, leading to greater accuracy and fairness of marking. The tutor-entered feedback on the hand-in and the general comments were considered to be a good feature of the modified version. The external marker considered the system to be extremely useful and made several highly supportive comments related to the system. The quality and quantity of feedback provided was considered to be much better in the second prototype than in the first.

6.1 Suggested improvements

Based on the findings from the evaluation of the second stage prototype it was decided to produce a third version with new modifications suggested by the markers and external moderator. It was suggested that tutors should be able to add to or modify the automated feedback at each stage of marking as well as adding general comments at the end. It was further suggested that these additional comments be saved to a database file so that they might be used again later. In addition an option to show an image of the student was suggested, although this suggestion did not receive universal support from all present as some favoured more anonymous marking.

7. Development of the third prototype

In order to allow for the comments from markers and the external moderator a third stage prototype is currently being developed. In this version several improvements are being made.

A more robust and secure database system is being employed which will not only store the automated feedback for each question section but will also store additional feedback comments that are added to each question to supplement the automated feedback. A list of these is then available when the question is again marked for another learner and can be added to the feedback with a single click.

Developing a high quality feedback database is an important part of the system and at present this is a time consuming and demanding task. In order to simplify this task, the system is being modified in order to make it simpler for the tutor to enter their marking scheme and feedback statements and create a feedback file that can be used by the system. At present the teacher must create a text file containing a great deal of meta-information relating to the format of the questions along with the feedback for each option. In the new system the tutor need only enter the marking scheme and the feedback. The feedback database file is then formatted by the system.

An option to display an image of each student has also been added, although this may be switched on or off. It was also suggested that it be possible to modify marks without needing to re-enter the complete set of marks for a student and this feature was added. Finally the feedback distribution has been improved so that it is no longer necessary to use a word processor to distribute the feedback via electronic mail merge. This function is now achieved from within the application directly. Figure four shows an example screen of the latest prototype. In this version the user interface has been improved. Teachers are able to set up assessments and feedback much more easily.
The systems described in the previous sections were used to provide feedback for more than 800 students. Table five shows the breakdown of the students, levels and modules for which the application was employed.

Table 5: Students, levels and modules and assignment details for which the systems were used

<table>
<thead>
<tr>
<th>Module</th>
<th>Number</th>
<th>Level</th>
<th>Assignment type</th>
</tr>
</thead>
<tbody>
<tr>
<td>E Media Design</td>
<td>320</td>
<td>First year BSc</td>
<td>Practical</td>
</tr>
<tr>
<td>Games Development</td>
<td>180</td>
<td>Second year BSc</td>
<td>Practical and theory</td>
</tr>
<tr>
<td>New Entertainment systems</td>
<td>160</td>
<td>Third Year BSc</td>
<td>Practical and theory</td>
</tr>
<tr>
<td>Multimedia System Design</td>
<td>140</td>
<td>MSc</td>
<td>Theory</td>
</tr>
<tr>
<td>Referred Deferred coursework</td>
<td>50</td>
<td>All levels</td>
<td>Practical and theory</td>
</tr>
</tbody>
</table>
The software was used for a fairly wide range of learners, from first year through to master’s level for a range of practical and theory assignments. More than 850 students received feedback from the system. Data from student feedback questionnaires suggested that the feedback delivered was of good quality. Data was only available for approximately half of the modules involved in the study. This data took the form of written comments. In general comments were complimentary about the system. Of 24 comments related to the feedback provided only 1 was slightly negative.

“Although the feedback was quick I thought I did better than the grade I got”. (Student level 1)

This was a general problem. The more detailed the feedback provided the more likely students were to question their grades. Most students provided a more positive view of the feedback they were provided with as the following example shows.

“… feedback was really fast and delivery by email is a great idea. I got this 3 weeks after. On other modules we get marks months after hand in”. (Student level 3).

The majority of comments from students were similar to that of the level 3 student above. In general it suggested that the application functioned well and the feedback provided was good and appropriate.

8. Use of the modified system

Prototype version four has been developed, however it has experienced several problems and despite having additional functionality it is not robust enough for reliable use. Further development work is required on this system before it can be used generally. However, based on the positive results obtained from the evaluation of the use of the third prototype, it was decided to extend the use of this system to include more modules at BSc and MSc level and hence more learners. In all, more than one thousand additional assessments have been undertaken with the third iteration prototype so far this year. An important addition was the use of the prototype to assess group project work at master’s level. Previously such work had proven difficulty to assess and difficult to provide learners with good and timely feedback because of the complexity of marking.

The module had four summative assignments. The first was an individual online multiple choice test covering the principles of multimedia design. The work leading up to the first assignment was intended to prepare learners for the latter three assignments. The 55 student on the module then split into 22 groups of between 2 and 4 persons. The second group assignment related to the development of a prototype Flash website in which students, in groups, produce a minimal content software prototype: essentially the basic structure of the website together with an animation to promote it. This was submitted along with documentation relating to the website’s information architecture, the goals and mission of the site, grouping and labelling of content, tree structure diagrams and design ideas for the visual appearance of the site. The third assignment related to students’ evaluations of other groups’ websites and finally, in the fourth assignment, the groups reform and redevelop their website based on the feedback obtained both from the tutors and from the other students on the module. The feedback they received from their work on the second assignment (the minimal content website) was therefore extremely important as this was used to guide the development of the full content website for assignment four. An example of the feedback provided for one part of the assignment is shown in table 6 below.

Table 6: Example of feedback comments related to marks awarded for one section of the documentation

<table>
<thead>
<tr>
<th>Mark awarded</th>
<th>Feedback comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>The goal and mission is superbly clear, unique and specific. Scenarios describe in depth credible potential users with credible search strategies. Analysis of competitor sites is deep. This will be of industrial standard.</td>
</tr>
<tr>
<td>4</td>
<td>Very good goal and mission. Good scenarios and good analyses of competitor sites. This will be high standard.</td>
</tr>
<tr>
<td>3</td>
<td>Unimaginative but credible goal and mission. Scenarios a bit generic but with some local colour. Analysis of competitor sites not great but generally correct.</td>
</tr>
<tr>
<td>2</td>
<td>Uninteresting goal and mission. No detail in the scenarios – or ones taken from a very post facto point of view. Analysis of competitor sites done only cursorily.</td>
</tr>
<tr>
<td>1</td>
<td>Uninteresting goal and mission. Scenario not credible or not there. Analysis of other sites purely descriptive without analysis</td>
</tr>
</tbody>
</table>
It was interesting to compare group performance on assignments 2 and 4. It would be expected that performance on assignment 4 would be likely to improve if feedback on assignment 2 were effective. Table 7 shows a comparison of the marks for assignment 2 and 4 on this module.

**Table 7**: A comparison between the marks for group working on two related assignments, the second test (Ass.4) took place after feedback was provided on Assignment2

<table>
<thead>
<tr>
<th>Group</th>
<th>Assignment 2</th>
<th>Ass. 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 01</td>
<td>26</td>
<td>26.5</td>
</tr>
<tr>
<td>Group 02</td>
<td>37</td>
<td>36.5</td>
</tr>
<tr>
<td>Group 03</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>Group 04</td>
<td>30</td>
<td>22</td>
</tr>
<tr>
<td>Group 05</td>
<td>29</td>
<td>35</td>
</tr>
<tr>
<td>Group 06</td>
<td>31</td>
<td>32.5</td>
</tr>
<tr>
<td>Group 07</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td>Group 08</td>
<td>38.5</td>
<td>42</td>
</tr>
<tr>
<td>Group 09</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Group 10</td>
<td>28.5</td>
<td>35</td>
</tr>
<tr>
<td>Group 11</td>
<td>36</td>
<td>29.5</td>
</tr>
<tr>
<td>Group 12</td>
<td>28.5</td>
<td>38</td>
</tr>
<tr>
<td>Group 13</td>
<td>28</td>
<td>44</td>
</tr>
<tr>
<td>Group 14</td>
<td>45.5</td>
<td>13</td>
</tr>
<tr>
<td>Group 15</td>
<td>26.5</td>
<td>28.5</td>
</tr>
<tr>
<td>Group 16</td>
<td>46</td>
<td>38.5</td>
</tr>
<tr>
<td>Group 17</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>Group 18</td>
<td>29.5</td>
<td>24.5</td>
</tr>
<tr>
<td>Group 20</td>
<td>18</td>
<td>31</td>
</tr>
<tr>
<td>Group 22</td>
<td>22</td>
<td>35.5</td>
</tr>
<tr>
<td>Group 23</td>
<td>38.5</td>
<td>26</td>
</tr>
<tr>
<td>Group 24</td>
<td>32</td>
<td>30</td>
</tr>
<tr>
<td>MEAN</td>
<td>30.4</td>
<td>30.8</td>
</tr>
</tbody>
</table>
Statistical analysis of the results shown in table 7 using an ANOVA showed no significant difference between the performance on the two assignments (p<0.4). Fourteen groups scored higher on the second assignment while eight groups scored lower. There was anecdotal evidence that some features of the assignment (the website design in particular) were better than in previous years. The results of this are hard to interpret however as multiple variables were involved. The assignment was a group assignment and group composition was a likely source of variance for example. Although it was hard to draw conclusions from this result it offers interesting possibilities for future studies.

Student comments on the nature of the feedback they received were elicited in end of module questionnaires. A selection of comments received is shown in table 8 below.

**Table 8: Examples of student comments on the feedback received.**

<table>
<thead>
<tr>
<th>Student</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>I got the marks back in less than a month after submitting. The feedback was very quick and email idea was good. Better than discussing the assignment in a lecture months later.</td>
</tr>
<tr>
<td>02</td>
<td>Comments were helpful</td>
</tr>
<tr>
<td>03</td>
<td>I would like all my feedback done by email.</td>
</tr>
<tr>
<td>04</td>
<td>The feedback document was to the point and useful to me. I could see where I went wrong and how to improve for the next assignment.</td>
</tr>
<tr>
<td>05</td>
<td>It was clear to me how to improve next time.</td>
</tr>
<tr>
<td>06</td>
<td>I did not agree with everything that was written but it was very detailed and I could argue with the lecturer about my score.</td>
</tr>
<tr>
<td>07</td>
<td>I prefer to discuss my results person to person. I suppose this is the next best thing in such a big group. It was very fast compared to other modules and it was fair.</td>
</tr>
<tr>
<td>08</td>
<td>I don’t read my uni email regularly so I missed my grade and the feedback till later. It would be good if it could be sent to my hotmail.</td>
</tr>
<tr>
<td>09</td>
<td>The feedback was simple and easy to understand. It was detailed so I could see exactly what I did right and where I went wrong.</td>
</tr>
<tr>
<td>10</td>
<td>There was quite a lot to read but at least I have a record of it with my marks.</td>
</tr>
</tbody>
</table>

A summary of the large number of comments received is shown below.

- Speed of automated feedback system was valued
- Feedback comments were in general helpful to students and of the right level of detail
- Students were able to argue about and discuss the marks awarded as the feedback was detailed enough to allow good checking (some tutors were unhappy with this).
- Feedback comments were cited as leading to improvement in work and good for revision
- Feedback provided in lectures was too general and was less valuable
- Feedback was simple and easy to understand
- Provided a permanent record which other formats did not
- Some students reported that there was lots to read; Some students evidently don’t like reading
- Some preferred face to face feedback
- Email delivery was sometimes missed or spammed.
- Students didn’t always agree with feedback.
9. Discussion

The research described here has presented the development, testing and evaluation of an automated feedback and marking system. The system was shown to be efficient and useful to both students and staff using the system. The evaluation of the system by staff and students suggests that the feedback quality was good and was delivered quickly and effectively using electronic mail. This was a vast improvement when compared to the manual methods used previously. The quality of feedback is vital and this was improved through modifications to the system, using the best features of automation and also by allowing tutors to add and save their own comments for later use. The use of a user-centred iterative prototyping approach involving staff and students was vital to the development of the system.

It was important that the system developer was able to understand the detailed requirements and functions of the system based upon the thoughts and opinions of a range of users. In this way the system was more likely to be accepted by colleagues and external examiners and more likely to be beneficial to all. Currently the third stage prototype is being tested prior to use with learners. It is hoped that the next version will, in addition to practical and mixed practical/theory examinations, be used in a pure essay type test, where it is expected to be especially beneficial. It is hoped that in later stage prototypes feedback comments will be inserted directly into the documents being assessed at an appropriate place in the text. This idea is currently under consideration for prototype version four.

References


Abstract: The correct grading of free text answers to exam questions during an assessment process is time consuming and subject to fluctuations in the application of evaluation criteria, particularly when the number of answers is high (in the hundreds). In consequence of these fluctuations, inherent to human nature, and largely determined by emotional factors difficult to mitigate, it is natural that small discrepancies arise in the ratings assigned to similar responses. This means that two answers with similar quality may get a different grade which may generate inequities in the assessment process. Reducing the time required by the assessment process on one hand, and grouping the answers in homogenous groups, on the other hand, are the main motivations for developing the work presented here. We believe that it is possible to reduce unintentional inequities during an assessment process of free text answers by applying text mining techniques, in particular, automatic text classification, enabling to group answers in homogeneous sets comprising answers with uniform quality. Thus, instead of grading answers in random order, the teacher may assess similar answers in sequence, one after the other. The teacher may also choose, for example, to grade free text answers in decreasing order of quality, the best first, or in ascending order of quality, starting to grade the group of the worst answers. The active learning techniques we are applying throughout the grading process generate intermediary models to automatically organize the answers still not fixed in homogeneous groups. These techniques contribute to reduce the time required for the assessment process, to reduce the occurrence of grading errors and improve detection of plagiarism.

Keywords: text mining, active learning, free-text assisted grading

1. Introduction

The assessment of free text answers is a demanding process requiring for a great effort from the evaluators, particularly when the number of answers to assess is high – in the hundreds. This process demands for high concentration levels and for long periods of time leading to fluctuations in the evaluator’s level of concentration and mood.

Moreover, evaluating answers in random order or, at least, without any order related to the quality of the answer itself can lead to situations in which different grades can be assigned to two answers of similar quality. Evaluating an answer of average quality after evaluating a set of answers of much lower quality may lead the evaluator to inflate the grade assigned to the average answer. Similarly, when that same average answer is evaluated after evaluating a number of very high quality answers, may trigger the opposite effect which will result in grading the average answer lower than if it was assessed in other circumstances. Two answers of similar quality, assessed after a series of high quality and low quality answers, will get different grades with some probability. These effects may be more evident when the evaluation process is longer.

These effects could be reduced if the selection of the next answers to access is made judiciously – to have the answers of similar quality being evaluated sequentially – and by ascending or descending order of quality – to avoid sharp fluctuations in the emotional context and the mood of the evaluator.

We believe it is possible to standardize the application of assessment criteria for evaluating free text answers by applying text mining techniques. In particular, using automatic text classification, to identify the different levels of answers’ quality and to group answers according to those levels, may reduce mood fluctuation in the evaluators. These techniques allow (a) reduce the time needed to complete the evaluation process (b) reduce the occurrence of errors of evaluation and (c) better detect cases of plagiarism.
Our current work applies active learning techniques and automatic text classification to organize free text answers to exam questions in homogeneous groups and then to sort them according to their quality.

This process requires the evaluator (user, in general) to provide a pre-labeled set of answers based on which an automatic classification model will be built. The labels or classes represent the grades to assign to every answer. To generate these classification models we rely on Support Vector Machines (SVM) with a linear kernel. SVM (Chapelle et al., 2006) classifiers are recognized as one of the most suited for text classification tasks.

This classification model is then applied to all answers that have not yet been assessed assigning an automatic label (rate) to them. The grouping of responses in homogeneous groups and their ranking will be based on these provisional ratings assigned automatically.

To generate a classification model that can recognize all relevant rates we need to obtain pre-labeled answers from each of these groups. These are previously labeled by the user. This initial phase – until you can obtain a classification model that covers all rates of relevance and that is sufficiently accurate – is more demanding and more prone to assessment errors. It is therefore important to make this a short-term phase, requiring the evaluator to evaluate only a small number of answers until all grades have been identified and characterized through these pre-labeled examples. To minimize the effort required for this initial phase and, simultaneously, shorten this learning phase we use a strategy of active learning called D-Confidence (Escudeiro et al., 2009).

Active learning techniques select the most informative answers, given the known evidence, avoiding asking the evaluator to rank responses from low value added. This is how active learning reduces the workload required from the evaluator to build an appropriate classification model. Experimental results show that it is possible to obtain valid classification models and improve the assessment processes.

The remaining sections of this paper present a brief review of the area of active learning, in Section 2 and describe the D-Confidence algorithm in Section 3. Section 4 describes our proposal for assisted rating of free text answers. In Section 5 we present the evaluation that has been performed and Section 6 states our conclusions and future work.

2. Active learning

There are several paradigms suited to automatic classification with numerous applications.

The paradigm of **supervised learning** allows you to specify arbitrary concepts, specific to a given problem. However, it requires a set of fully labeled data which can be prohibitive when tagging cases is a costly process, as is usually the case with text documents. In addition, it demands for exemplary instances, previously labeled, for all classes to learn.

The **semi-supervised** classification (Chapelle et al., 2006) allows the specification of particular needs without requiring a prior process of intensive labeling. Nevertheless, it also requires a minimum set of pre-labeled cases covering all the classes to learn.

**Unsupervised** algorithms require no prior labeling, however, this paradigm does not give any chance for the user to guide the model towards specific needs and there is no a priori guarantee that the groups automatically generated without any input from the user are aligned with the groups of interest to the user and the current problem at hand.

**Active learning** techniques which seem more suited to our purposes select the next case to label wisely, prompting the user for its label. The most informative case will be selected by the learning algorithm rather than being randomly selected as in the case in supervised learning. It is expected that a lower number of labels is required to achieve the same accuracy when compared to the fully supervised learning setting.

Active learning approaches (Angluin, 1988; Cohn et al., 1994; Muslea et al., 2006) reduce label complexity – the number of queries that are necessary and sufficient to learn a concept – by analyzing unlabeled cases and selecting the most useful ones once labeled. Queries may be artificially generated (Baum, 1991) – the query construction paradigm – or selected from a pool (Cohn
et al., 1990) or a stream of data – the query filtering paradigm. Our current work is developed under the query filtering approach.

The general idea in active learning is to estimate the value of labeling one unlabeled case. Query-By-Committee (Seung et al., 1992), for example, uses a set of classifiers – the committee – to identify the case with the highest disagreement. Schon et al. (2000) worked on active learning for Support Vector Machines (SVM) selecting queries – cases to be labeled – by their proximity to the dividing hyperplane. Their results are, in some cases, better than if all available data is used to train. Cohn et al. (1996) describe an optimal solution for pool-based active learning that selects the case that once labeled and added to the training set, produces the minimum expected error. This approach, however, requires high computational effort. Previous active learning approaches (providing non-optimal solutions) aim at reducing uncertainty by selecting the next query as the unlabeled example on which the classifier is less confident (Lewis and Gale, 1994).

Batch mode active learning – selecting a batch of queries instead of a single one before retraining – is useful when computational time for training is critical. Brinker (2003) proposes a selection strategy, tailored for SVM, that combines closeness to the dividing hyperplane – assuring a reduction in the version space close to one half – with diversity among selected cases – assuring that newly added examples provide additional reduction of version space. Hoi et al. (2006) suggest a new batch mode active learning relying on the Fisher information matrix to ensure small redundancy among selected cases. Li et al. (2006) compute diversity within selected cases from their conditional error.

Dasgupta (2005) defines theoretical bounds showing that active learning has exponentially smaller label complexity than supervised learning under some particular and restrictive constraints. This work is extended in Kaariainen (2006) by relaxing some of these constraints. An important conclusion of this work is that the gains of active learning are much more evident in the initial phase of the learning process, after which these gains degrade and the speed of learning drops to that of passive learning.

Agnostic Active learning (Balcan et al., 2006), $\mathcal{A}^2$, achieves an exponential improvement over the usual sample complexity of supervised learning in the presence of arbitrary forms of noise. This model is studied by Hanneke (2007) setting general bounds on label complexity.

All these approaches assume that we have an initial labeled set covering all the classes of interest.

Clustering has also been explored to provide an initial structure to data or to suggest valuable queries. Adami et al. (2005) merge clustering and oracle labeling to bootstrap a predefined hierarchy of classes. Although the original clusters provide some structure to the input, this approach still demands for a high validation effort, especially when these clusters are not aligned with class labels.

Dasgupta et al. (2008) propose a cluster-based method that consistently improves label complexity over supervised learning. Their method detects and exploits clusters that are loosely aligned with class labels.

Among other paradigms, it is common that active learning methods select the queries which are closest to the decision boundary of the current classifier. These methods focus on improving the decision functions for the classes that are already known, i.e., those having labeled cases present in the training set. The work presented in this paper diverts classifier attention to other regions increasing the chances of finding new labels.

In this work, we used the active learning algorithm called D-Confidence. This approach affects the learning process and labeling diverting it to regions of space where not yet explored. Thus minimizing the number of questions you need to do to find cases representing all classes.

The D-Confidence selects cases based on a tagging feature that adds confidence that the current classifier has a particular class - a traditional criterion in active learning - the distance between this case and the previously known cases of that class by the classifier.
This criterion is biased for cases that do not belong to known classes - lower confidence - and that are located in regions of space where unexplored - Distance high to known classes. This way is more efficient in the even coverage of the concepts to learn.

Common techniques of active learning based on a low confidence of the classifier instantiated to select cases to label (Angluin 1988) and assume that the chaos pre-labeled cover all the classes to learn - this assumption is not valid in our case. These approaches use the classification model from each iteration to calculate the confidence in each class for each unlabeled case, then selecting the unlabeled case with the least confidence.

Other more recent approaches make use of unsupervised classification to find cases to label (Dasgupta et.al, 2008).

3. D-Confidence

The most common active learning approaches rely on classifier confidence to select queries (Angluin, 1988), assume that the pre-labeled set covers all the labels to learn and are focused on accuracy. Our scenario is somehow different: we do not assume that we have labeled cases for all classes and, besides accuracy, we are mainly concerned with the fast identification of representative cases from all classes. To achieve our goals we rely on the d-Confidence selection criterion (Escudeiro et al., 2009), which is effective in the early detection of exemplary cases from unseen classes. Instead of relying exclusively on classifier confidence we propose to select queries based on the ratio between classifier confidence and the distance to known classes. D-Confidence, weights the confidence of the classifier with the inverse of the distance between the case at hand and previously known classes.

D-Confidence is expected to favor a faster coverage of case space, exhibiting a tendency to explore unseen regions in case space. As a consequence, it provides faster convergence than confidence alone. This drift towards unexplored regions and unknown classes is achieved by selecting the case with the lowest d-Confidence as the next query. Lowest d-Confidence is achieved by combining low confidence – probably indicating cases from unknown classes – with high distance to known classes – pointing to unseen regions in the case space. This effect produces significant differences in the behavior of the learning process. Common active learners focus on the uncertainty region asking queries that are expected to narrow it down. The issue is that the uncertainty region is determined by the labels we known at a given iteration. Focusing our search for queries exclusively on this region, while we are still looking for exemplary cases on some labels that are not yet known, is not effective.

Unknown classes hardly come by unless, by chance, they are represented in the current uncertainty region.

In active learning, the learner is allowed to ask an oracle (typically a human) to label examples – these requests are called queries. The most informative queries, given the goals of the classification task, are selected by the learning algorithm instead of being randomly selected as is the case in passive supervised learning. The general idea in active learning is to estimate the value of labeling one unlabeled case. The general algorithm below is the basement for active learning classification.

```
Initialize U_0, L_0
Train the learner in L_0 to generate h_0
Run h_0 to classify instances in U_0
i = 1
While not stopping criteria {
    Select q_i = <x_i> pertencentea U_{i-1}, the most adequate queries from U_{i-1}
    Ask the oracle for their labels, c_i
    L_i = L_i U <x_i, c_i>
    U_i = U_i minus <x_i, c_i>
    Train the learner in L_i to generate h_i
    Run h_i to classify instances in U_i
    i++
}
```
D-Confidence is a particular algorithm to select queries which is particularly suited to reduce labeling effort at an initial stage of the classification process. This initial stage, when the user is still trying to find representative cases from all the grades he is interested in, is the most critical in our current work.

D-Confidence is based on the ratio between confidence and distance among cases and known classes (Eq. 1).

\[
\text{Eq.1} \quad dConf(u) = \max_k \left( \frac{\text{conf}_k \mid u}{\text{median dist}(u, xlab_k)} \right)
\]

For a given unlabeled case, \( u \), the classifier generates the posterior confidence w.r.t. known classes. Confidence is then divided by an indicator of the distance between the unlabeled case at hand and all labeled cases belonging to class \( k \), \( xlab_k \). This distance indicator is the median of the distances between the unlabeled case at hand and all labeled cases belonging to the class. We expect the median to soften the effect of outliers. D-Confidence for each known class is computed by dividing class confidence for a given case by the aggregated distance of the unlabeled case to that class. Finally, we compute d-Confidence of case \( u \), \( dConf(u) \), as the maximum d-Confidence on individual classes.

4. Assisted grading of free text answers

To evaluate the gains of our approach, we have developed a prototype (Figure 1) for assisted assessment and grading of free text answers to examinations questions.

![Figure 1: Learning process, evaluation and grading of answers to train the automatic classification model](image-url)

This prototype compiles the free-text answers to a given question and pre-processes them into a model that is suited for the application of automatic classification techniques (pre-processing). In particular, the set of answers to a given question is converted into a TFxIDF matrix (Weiss 2005 et.)
al) that represents, each answer, a text document, by a row in this matrix. The columns, in turn, represent the terms or, more generally, the elements of natural language that occur in the answers. The techniques of automatic text classification are then applied to this array of vectors representing the corpus of the students’ answers to a given question.

The assessment process is initiated by the user (teacher) that analyzes and classifies a set of answers (a minimum of two is required). These will be the seed labeled cases required to kick off d-Confidence. In our prototype, the ratings given by the teacher should be referred to the scale 0-100. These answers, previously labeled by the teacher, are used to train a SVM classifier (the learning task). This classifier generates a classification model for a set of labels, or classes, representing grades. The maximum number of different grades is automatically set by the prototype based on the number of answers in the set being assessed. This limitation stems from the fact that we need a minimum number of examples for each class so that we can ensure some accuracy in automatic classification.

Depending on the number of answers, \( N \), to a certain question, we set the maximum number of classes (grades), \( K \), that the classifier is able to learn using Eq.2:

\[
K = \max\left(3, \frac{N}{10}\right)
\]

The number of classes to learn will always be between 1 and \( K \). When assessing students’ answers, if the teacher assigns a grade of \( C \) (in the scale 0-100) then, for the purpose of organizing the homogeneous groups, this answer will be assigned to class \( k_i \) – the corresponding grade to \( C \) that our classifier can distinguish with a certain confidence. Equation 3, is used to define the class to which an answer graded by the teacher with a mark of \( C \) belongs to:

\[
k_i = \text{int}\left(\frac{C \times K}{100}\right) + 1
\]

Thus, in a case with 25 answers to evaluate, for instance, we will have them grouped into three groups or classes/grades. The first class groups students’ answers rated by the teacher with a score between 0% and 33%, class two, groups the answers graded between 34% and 66% and class three, groups the answers rated between 67% and 100%.

Once the number of different classes mapping teacher grades is set and the first classifier is generated, from a set of pre-labeled answers, the automatic learning process runs in parallel with the grading process until the moment when the teacher finds that the automatic classification model is reasonable.

Every iteration of the learning process begins by applying the current classification model to all the answers that have not yet assessed. This model assigns a class/grade to each un-assessed answer with a given confidence.

After these classes are assigned to each answer the D-Confidence active learning algorithm is applied to sort the unlabeled answers and to present them to the teacher in descending order of their informative value to the current classification model, i.e., by increasing order of their d-Confidence measure.

The teacher evaluates the answer with greater informative value, which is then added to the set of all the answers previously evaluated. This set of teacher-labeled answers is then used to generate a new classification model – which is expected to be more accurate because it was trained on the basis of more cases properly labeled – and the process iterates.

These iterations are repeated until the classification model can estimate grades for un-assessed answers in a manner deemed sufficiently accurate by the teacher. Thereafter, this lastly generated classification model is applied to group the answers that have not been assessed yet in groups of
homogeneous quality joining together the answers that have been assigned to the same grade by the classification model.

These homogeneous groups are then presented to the teacher by ascending or descending order of their quality according to the preference of the teacher.

5. Evaluation

The evaluation of our proposal was based on a real dataset, with 31 free text students’ answers to a question from an examination of the course of Software Engineering.

As the number of responses is 31 we have considered three classes corresponding to the grades in the ranges 0-33%, 34-66% and 67-100%, according to Eq.2 and Eq.3.

All answers have been assessed by the teacher that has assigned a grade between 0 and 100 to every answer. These teacher grades are hidden from the classifier. They will be used for evaluation purposes only to compare automatic grades to those assigned by the teacher and so compute the accuracy of the automatic classifier.

The accuracy of the classifier is calculated on the set of answers that were not used to train the current classifier.

In the first iteration only three quotations from different classes where provided to the classifier. The classification model generated from these three labeled cases was applied to the remaining 28 student's answers (those that have not been used for training) assigning a grade to each one. From these classes automatically assigned by the classifier, 6 correspond to the true class of the answer, as assessed by the teacher, which corresponds to an accuracy of 21% (6/28). In this first experiment, the classifier was trained based on three cases and does not have an acceptable accuracy. Confidence is low and the teacher may not accept the suggestions given by the classifier.

In a second iteration 6 labeled answers were provided to the classifier, two for each different grade. The percentage of correct predictions in all the answers that were not used to train the classifier is 36% (9 hits in 25). In this iteration the confidence index is still low, and the teacher may not accept the suggestions of the classifier.

In the third iteration, nine labels were submitted, three labeled answers from each grade. The percentage of correct predictions is 45% (10/22).

Finally, in a fourth iteration we have provided 12 graded answers to the classifier, four from each grade. The percentage of correct predictions is now 68% (13/19).

Although an accuracy of 68% is not a high standard, it seems reasonable for the intended purpose. We must recall that the purpose of automatic classification in our current work is not to make the automatic attribution of grades to the students’ answers, but simply to allow grouping answers in homogeneous groups to reduce errors in the assessment process and to reduce the time required by the assessment process.

6. Conclusions

Grouping responses of similar quality, those that are worth similar grades, prior to having them assessed by the teacher, may reduce errors arising from a random assessment order constantly interspersing answers with high quality with others of lesser quality.

For the assessment process to be done this way it is necessary to have some semi-automatic way of creating these homogeneous groups and of sorting them by ascending or descending order of their quality. This stems from the effort that, otherwise, would necessarily be required from the teacher to organize students’ answers in these groups, probably would make the whole process useless.

Our evaluation, although not too extended, provides interesting results and indicates that, despite a poor accuracy, it is possible to create homogeneous groups of answers using techniques of automatic text classification.
One reason for the low accuracy probably relates to the fact that the answers are all on the same topic. This, being essential for such an application, reduces the usefulness of the classifier. A test made with a corpus on different topics has achieved 96% of correct suggestions (the suggestion referred to the theme that the document concerned). Under these circumstances (the need to distinguishing text documents all referring to the same topic) we are studying the potential impact on the accuracy of the classifier of the use of other models of representation of the answers besides TFxIDF. We are examining the benefits of using models based on multi-term attributes (N-grams), models focused in certain categories of words (part-of-speech tagging) and string kernels (Lodhi 2002).

References

Enhanced Automatic Question Creator – EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e-Education

Christian Gütl1, 2, Klaus Lankmayr1, Joachim Weinhofer1 and Margit Höfler1
1Institute for Information Systems and New Media (IICM), TU-Graz, Austria
2Curtin University, Perth, Australia

Abstract: Research in automated creation of test items for assessment purposes became increasingly important during the recent years. Due to automatic question creation it is possible to support personalized and self-directed learning activities by preparing appropriate and individualized test items quite easily with relatively little effort or even fully automatically. In this paper, which is an extended version of the conference paper of Gütl, Lankmayr and Weinhofer (2010), we present our most recent work on the automated creation of different types of test items. More precisely, we describe the design and the development of the Enhanced Automatic Question Creator (EAQC) which extracts most important concepts out of textual learning content and creates single choice, multiple-choice, completion exercises and open ended questions on the basis of these concepts. Our approach combines statistical, structural and semantic methods of natural language processing as well as a rule-based AI solution for concept extraction and test item creation. The prototype is designed in a flexible way to support easy changes or improvements of the above mentioned methods. EAQC is designed to deal with multilingual learning material and in its recent version English and German content is supported. Furthermore, we discuss the usage of the EAQC from the users’ viewpoint and also present first results of an evaluation study in which students were asked to evaluate the relevance of the extracted concepts and the quality of the created test items. Results of this study showed that the concepts extracted and questions created by the EAQC were indeed relevant with respect to the learning content. Also the level of the questions and the provided answers were appropriate. Regarding the terminology of the questions and the selection of the distractors, which had been criticized most during the evaluation study, we discuss some aspects that could be considered in the future in order to enhance the automatic generation of questions. Nevertheless the results are promising and suggest that the quality of the automatically extracted concepts and created test items is comparable to human generated ones.

Keywords: e-assessment, automated test item creation, distance learning, self-directed learning, natural language processing, computer-based assessment

1. Introduction

Highest flexibility is required from the members of our modern world in terms of continuous adaptation of knowledge and skills. Formal education in primary and secondary settings but even academic settings is not sufficient any more for our ever-changing and knowledge-driven society. Thus life-long learning is the key in such an environment and new pedagogical approaches such as exemplary-based learning and self-directed learning are becoming increasingly popular. (Gütl, 2010) Commonly agreed and widely discussed in literature, such as in Bransford, Brown and Cocking (2000), assessment has not only be seen as an integrated part of the learning processes but also feedback to students and teachers is important to adapt the learning process and improve the learning outcome. Assessment activities are resource intensive and time-consuming which has motivated different computer-supported and computer-assisted approaches. The various approaches range from applications supporting human-based marking and feedback to applications, which support automated assessment. E-assessment tools can certainly reduce effort and improve feedback, however, the creation of appropriate test items is a time consuming task, in particular to assess content alternatives and different knowledge levels in adaptive e-learning environments. Moreover, in self-directed learning settings or more general in life-long learning settings there is no pre-defined learning content and students can select content from open or closed repositories or even Web content. Consequently it is almost impossible to provide prepared test items for such kind of learning. (Gütl, 2008)
mated approaches. One important research strand in this context is semi-automated and fully-automated test item creation. A first simple solution has combined an approach for statistic text summaries and a named entity detection algorithm (Gütl, 2008). Findings of the first approach have led to an enhanced approach combining statistical, structural and semantic analysis for concept detection, and based on that different types of test items have been created (Gütl, Lankmayr, & Weinhofer, 2010). First pilot trials, a user study and findings from the development point of view have resulted in further improvements of the prototype.

In this paper, which is an extended version of the conference paper of Gütl, Lankmayr and Weinhofer (2010), we want to outline the enhanced version of the prototype and report about the most relevant finding of a user study focusing on the perception of the quality of the automatically created test items. To this end, the paper is structured as followed: first we will give background information and related work on both the concept extraction and automatic test item creation. This is followed by requirements, design and development of the enhanced prototype, the Enhanced Automatic Question Creator (EAQC). A discussion from the users’ viewpoint as well as user study of the quality of the extracted concepts and created test items give first insights of the practical usage.

2. Background and related work

Following the basic idea of the proposed approach of the automated creation of assessment items, one of the most important tasks is the identification of the most relevant concepts form of natural language texts of the learning content, which is an active research topic in past and present, such as in (Moens & Angheluta, 2003; Villalon & Calvo, 2009). A short overview of the historic developments of concept extraction is based on (Gütl et al, 2010; Weinhofer, 2010). Early and initial ideas of concept extraction can be based on research of Luhn who found statistic relationships of words in textual content (Luhn, 1957). In the late 1970s Edmundson improved this method by combining cue phrases, word frequencies, title words and the position of words in a paragraph. Kupiec, Pederson and Chen (1995) extended this method by considering acronyms and proper nouns additionally. Frank et al (1999) created a domain-specific key phrase extraction (KEA) that uses a Naive Bayes classification depending on word frequency and the position of the first occurrence of the word. KEA was extended by Turney (2003) who enhanced the algorithm by co-occurrences which consider the customariness of two words together in the WWW. Song, Han and Rim (2004) generated lexical chains and a concept score depending on word association, the depth in WordNet hierarchy and a semantic relation weight. Hassan, Mihalcea and Banea (2007) use a text rank algorithm that takes account of the context of a word by transforming the document into a graph and calculating node weights. Ledeneva, Gelbukh and García-Hernández (2008) evaluate n-grams, consisting of n words, instead of single words to determine the importance of concepts. A more detailed discussion of methods and approaches can be found elsewhere, such as at (Liu & Yang, 2009; Hovy, Kozareva & Riloff, 2009).

By further focusing on research of automated test item creation, an extensive literature review has shown just few pre-existing approaches and tools where most of the available tools support multiple choice items (Gütl, 2008; Lankmayr, 2010; Gütl et al, 2010). In an early and simple approach, Coniam (1997) identified the concept/expression by two distinctive ways: (a) user defined n-th word deletion depending on a predefined entry point, and (b) a part of speech tag. Distractors are extracted from a list derived from the Bank of England Corpus whereby these words have similar word frequencies in that corpus as the selected word. In the approach from Mitkov and Ha (2003), distractors based on given key terms are calculated by the use of WordNet. The questions are built by a rule based transforming of sentences into interrogative clauses. Machine learning was applied by Hoshino and Nakagawa (2005). Thus, k-Nearest neighborhood, naive Bayes classification and a suitable training set are utilized to identify the positions of the blanks in news articles for creating multiple choice items. Goto et al (2010) introduce a solution, which combines the following process steps: (a) extract appropriate sentences based on preference learning, (b) identify blank part based on conditional random field, and (c) create distractors based on statistical patterns of existing questions. Brown, Frishkoff and Eskenazi (2005) developed the REAP system, that is to provide texts suitable for users’ reading levels and to generate appropriate multiple choice but also assignment items. Some work can also be identified focusing on other test item types. By the help of WordNet using definitions, synonyms, antonyms, hyponyms and hypernyms question items are formed to improve word knowledge by evaluating user statistics. Chen, Liou and Chang (2006) have built grammar tests by transforming sentences extracted from the WWW. The transformation is done by applying manually generated patterns and is used for creating multiple choice items and error detection tests. Rus, Cai and Graesser (2007) introduced methods for generating questions with the help of patterns, templates and a special markup
language named QG-ML. The patterns are characterized by semantic, lexical and syntactical structures whereas the templates describe methods to implement these structures to generate questions. Heilman and Smith (2010) generate questions from reading materials by applying manually created rules and a ranking algorithm for items selection. Gütl (2008) described a system that uses automatic summary of a document to identify key concepts (named entities) and that generates completion tests as well as limited choice items.

The evaluation of the current state of research suggests that approaches using machine learning are strongly depending on the training set and the knowledge domain. Most of the illustrated systems are applying either statistical or semantic methods and are not able to fulfill the requests given by the variety of assessment item types. Moreover, pre-existing approaches and tools are not sufficiently flexible and extendable to support the above mentioned variety of application scenarios and learning settings. For this reason we developed a system, the Automatic Question Creator (AQC) as outline in Gütl et al (2010), which builds on a combination of statistical, semantic and structural analysis to accomplish a step-by-step extraction of relevant concepts from natural language texts. Insights of the first version of the prototype have led us to improve the system which is outlined in the subsequent sections.

3. Requirements, design and development

This section is an extended and updated version of the technical description outlined in Gütl et al (2010) and covers the technical aspects of the improved version of the automatic question creator tool, the Enhanced Automatic Question Creator (EAQC).

3.1 Objectives and high level requirements

Based on the findings and experiences of the first prototype development, the goal of the EAQC is to apply improved natural language processing methods which supports the creation of test items or even generates them automatically from the learning content of different languages. A flexible design should enable various groups to use the tool stand-alone or to integrate it in a learning platform as well as adjust the tool according to the specific learning setting. This has led us to specify to following requirements on an abstract level:

- Support of various input file formats from local file systems and from Internet resources
- Multilanguage support
- Domain knowledge and document structure independency
- Identification of most important concepts
- Creation of test items and reference answers based on identified concepts
- Support of open ended, single choice, multiple-choice and completion exercises
- Variability, configurability, modularity, extensibility and performance
- Interoperability with existing eLearning systems

3.2 Conceptual architecture and tools

The high-level conceptual design of the EAQC is outlined in Figure 1. It illustrates the core conceptual units and pre-existing tools as well. The system can be unfold into three main modules: (1) The Pre-processing module deals with format conversion of several file formats and online resources, text cleaning methods, language detection and transformation into an internal XML schema which contains all necessary data for further processing. In the current system English and German languages are supported, however, the flexible design easily enables to integrate other modules or tools to support other languages. (2) The Concept Extraction module performs structural, statistical and semantic analysis, runs term weighting and finally extracts the most suitable phrases; a detailed description is given in Section 3.3. (3) The Assessment Creation module determines the most appropriate sentence for each phrase and adds the previous and the following sentences to provide sufficient context information. Moreover the module identifies distractors and antonyms, creates question items and reference answers, and finally transforms those items in QTI standard.
The main components integrated in the implemented system are GATE and two lexical databases. The GATE framework, especially the ANNIE plug-in, is used for basic text processing and annotation. Thereby the text is split up into tokens and sentences, the part of speech classification as well as name entity recognition, noun chunking and co-reference resolution of each token are performed. (GATE, 2010) The semantic analysis is processed with WordNet in case of English language or GermaNet when performing analysis on a German text. (WordNet, 2010; GermaNet, 2009) Thereby semantic and lexical relations between words are calculated as well as distractors and antonyms are selected. Format conversion for Word, Open Document Text and HTML is utilized to transform the
input files into a HTML format by using JODConverter (2010). PDF files are transformed with the help of PDFBox (2010) that is able to extract the textual information from such files, structural information is added manually by applying predefined patterns. Content of the WWW, such as Wikipedia, is also supported as input source by the Automatic Question Creator. To ensure a high quality conversion especially to support Wikipedia content, a Wikipedia parser was implemented, to deal with the inconsistency of the provided HTML source code. Afterwards the generated HTML file is cleaned up using HTML Cleaner (2006) to ensure a conversion to XML with JDOM (2010). The concept extraction done by the EAQC is assisted by XtraK4Me of Schutz (2008) which was adapted to fit the requirements of the German language too. QTI exportation and rendering is done with JQTI (2008).

3.3 Data structure and applied methods

The main idea of our enhanced approach is to combine statistical, semantic and structural analysis to find most relevant words in learning content or more concrete concepts suitable for creating tests and exercises. Based on general word frequencies of the stemmed text the EAQC transforms those frequencies into weights for each word. In the second step of the process chain, these weights are adapted by a configurable set of algorithms that evaluate dependencies of the words according to the appearance in the text, such as in title, abstract, keywords, headlines. Also structure and formatting style as well as word types are considered in the process. Depending on the set of the highest weights and further configurable parameters the EAQC generates single choice, multiple-choice, completion exercises and open ended questions. Moreover the system is capable of exporting the test items including reference answers into the QTI format to allow integration into other learning and assessment systems.

In order to support the process chain, an internal data structure is applied which is organized into three main elements as illustrated in Figure 2. A **Word Element** contains all necessary textual and structural information of each token retrieved from GATE, WordNet and format conversion as well as from statistical and semantic analysis. According to the German language additional information is retrieved from GermaNet, the TreeTagger and the Durm German Lemmatizer. (TreeTagger, 1996; Durm, 2010) Each token is also associated with a **Weight Element** that stores a weight of each algorithm performed for concept extraction. The **Sentence Element** is calculated for each sentence in the text and contains the sentence boundaries, the related concepts and the sentence weight.

![Figure 2: Internal data representation](image)
The overall weight of words is composed of its statistical weight based on word occurrence \( w(1) \) (see Table 1, line 1) and several other weights \( w_i \) (see Table 1, line 2 - 11) that are retrieved by applying statistic, semantic and structural analysis. Most of these methods are subject to the distance of words in the used lexical databases hierarchies. The influence of each those weights on the overall weight can be adjusted by a set of independent parameters \( k_{m,1} \). Our first approach to the calculation of the overall weight \( w(1) \) for a word \( i \) is shown in equation (1), further experiments and improvements are subject to future work. To ensure stop word elimination only nouns and verbs are considered. A more detailed description of the weighting process and the applied methods can be found in Weinhofer (2010).

Table 1: Algorithms, weights and configurable parameters

<table>
<thead>
<tr>
<th>Module ( l_n )</th>
<th>Weight</th>
<th># Adjustable Parameters ( k_m )</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( w_{stat}(1) )</td>
<td>1</td>
<td>statistical weight, normalized number of occurrences of a stemmed word in a section</td>
</tr>
<tr>
<td>2</td>
<td>( w_{sim}(1) )</td>
<td>2</td>
<td>weight derived from statistical weights of similar words, depends on similarity measures retrieved from WordNet and GermaNet</td>
</tr>
<tr>
<td>3</td>
<td>( w_{title}(1) )</td>
<td>1</td>
<td>semantic relation to the words in the title</td>
</tr>
<tr>
<td>4</td>
<td>( w_{headline}(1) )</td>
<td>6</td>
<td>semantic relation to the words in the corresponding headline depending on the headline layer (up to 6)</td>
</tr>
<tr>
<td>5</td>
<td>( w_{abstract}(1) )</td>
<td>1</td>
<td>semantic relation to the words in the abstract</td>
</tr>
<tr>
<td>6</td>
<td>( w_{keywords}(1) )</td>
<td>1</td>
<td>semantic relation to keywords</td>
</tr>
<tr>
<td>7</td>
<td>( w_{annotation}(1) )</td>
<td>17</td>
<td>weight for the special annotations retrieved by GATE, the 17 annotation types can be handled individually</td>
</tr>
<tr>
<td>8</td>
<td>( w_{category}(1) )</td>
<td>25</td>
<td>weight according to the 25 unique beginners retrieved from WordNet and GermaNet</td>
</tr>
<tr>
<td>9</td>
<td>( w_{formatting}(1) )</td>
<td>1</td>
<td>Weight depending on the text formatting</td>
</tr>
<tr>
<td>10</td>
<td>( w_{keyphrase}(1) )</td>
<td>1</td>
<td>weight for phrases supplied form XtraK4Me algorithm</td>
</tr>
<tr>
<td>11</td>
<td>( w_{recursive}(1) )</td>
<td>2</td>
<td>recursive similarity weight calculation, consideration of lexical chains</td>
</tr>
</tbody>
</table>

In a further step, for each noun which is above a predefined threshold, a set of phrases that contain this word is built for each of the sections. Then all phrases of each set are weighted by summing up the overall weights of all words contained in a phrase. The highest weighted phrase of each set is chosen as potential concept. Finally the concept extraction is accomplished by building a collection of the best of these concepts for each section of the text.

\[
\begin{align*}
    w(i) &= w_{stat}(1) \cdot \left( k_1 + \sum_{l=2}^{11} \sum_{m} [(w_{l} \cdot k_{m}] \right) \\
&= w_{stat}(1) \cdot \left( k_1 + \sum_{l=2}^{11} \sum_{m} k_{m} \right)
\end{align*}
\]  

(1)

For Completion Exercises the previous and following sentences are added to the selected sentence to offer additional context information to the user. In all of those sentences the selected concepts get replaced with fill-in blank areas to avoid unnecessary hints. Multiple-choice item also requires distractor calculation. Basically the distractors are determined by searching coordinate terms for the whole
question phrase in WordNet respectively in GermaNet. If this calculation fails, the phrase gets split in all possible coherent n-grams and the coordinate terms for the longest sequence are randomly selected. In the worst case only a single word of a concept delivers suitable results. Due to the circumstance that there are very few proper nouns and no dates included in WordNet and GermaNet, a special case appears if the concept is assigned to a special annotation type. In this case three random phrases sharing the same annotation type are chosen as distractors from the underlying document. *Single choice items* can be generated by searching antonyms for single words in a concept and replacing the original word. Since the result of this procedure is seldom satisfying, the same method is repeated with all adjectives, verbs and nouns of the whole sentence. *Open ended exercises* are generated using several patterns depending on the special annotation type in the selected concept. Due to the fact of implementing a fully automatic assessment system the difficulty according to open ended questions is to compute a reference answer automatically. To meet this challenge the EAQC uses the text tiling algorithm to find the most proper text block containing the extracted concept.

The created test items are finally transformed into the QTI standard as single XML file for each question item. The reason for that exportation is to afford an opportunity of integrating the generated test items in learning management systems or other assessment tools. Currently a web service is developed to improve the flexibility in terms of submitting the learning content and to access the extracted phrases and the created test items.

4. Usage viewpoint

This section outlines EAQC from the user's point of view which is focused on the semi-automatic test item creation in a kind of interactive mode. The fully automated test item creation or batch mode processes the same steps but applies pre-configured settings. As the improvements of the enhanced tool (the EAQC) mainly have focused on methods of concept identification and test item creation, the graphical user interface has kept the same. Thus, the content in this section is a slightly adapted version of Gütl et al (2010) showing the process steps applied on the learning content of the case study (see also Section 5). The process steps are as follows: First, an input file in one of the supported formats has to be selected either from the local file system or from an Internet resource. The text is converted and filtered as well as a control output is generated. In the next step the user can induce the annotation process and the internal data structure is built. The result of the annotation is shown and the user can initiate the weighting process for concept identification. Figure 3 illustrates an example of a weighted text and the calculated weighting factors of a token which results from selected methods. In this step the user can initially set or change the weighting factors of the methods or even select and unselect methods to be applied (see Figure 4).

![Figure 3: Annotated and weighted text](image-url)
The next step in the process chain is the selection of the most important concepts to finally create the test items. Figure 5 illustrates a sample of concept extraction whereby the highest weighted phrase for each section of the content is listed on the screen. The user is enabled to deselect unwanted phrases as well as add unconsidered phrases or single words in a chapter of the text. Based on the final settings the test item creation is initiated. Different types of test items can be selected and instances of created items can be viewed. An example of a generated multiple-choice test item is outlined in Figure 6 that shows the representation of the QTI item in HTML.

Figure 4: EAQC configuration panel

Figure 5: Example of extracted concepts
5. Case study in academic education

To verify the implemented system, especially the quality of the extracted concepts and created test items, we conducted a study within the regular course “Information Research and Retrieval (ISR)” at Graz University of Technology at the end of the winter semester 2010/11. In particular, we were interested in how students evaluate automatically extracted concepts and test items (namely open-ended, single choice, multiple choice, and completion exercises, respectively) compared to concepts and test items generated by human.

5.1 Study setup

29 participants (4 female) took part in this study. They were 25.4 years on average (SD = 3.3), ranging from 22 to 39 years. Most of them (93.1%) were bachelor students; the rest were master students. Results from the tests delivered during the study (see below) were part of the final grading of the course, but note that the participation in the study was not a prerequisite for the completion of the course. All participants gave informed consent before attending the study. In order to generate questions with the EAQC, we modified a learning content (approximately 2,600 words) about “Natural Language Processing” (NLP) from Wikipedia (http://en.wikipedia.org/wiki/Natural_language_processing).

The procedure of the study was as follows: At the beginning, the scope and the time schedule of the study was briefly outlined by the experimentators. Participants were informed that they had to attend several learning activities during the session (see also Lankmayr, 2010, for a similar approach). The whole material (including the text, the instructions and all questionnaires) was presented as Web-based content. Furthermore, although almost all of the students were German-speaking, the learning content and the questionnaires were presented in English in order to enable comparing studies on international level. Participants were also asked to provide - if necessary - answers in English. After the introduction, students were asked to learn the text about “Natural Language Processing” (NLP) for 35 minutes and to briefly summarize it afterwards (Test 1; 10 minutes). Participants were not allowed to consult the given text during the test.

After a short break, the first of two main learning activities started. The goal of the first learning activity was that the students became familiar with the learning content. Similar to the operation method of the EAQC, students were asked to extract relevant concepts from the text first and to create eight test items (labeled as questions in the following) concerning the text afterwards. According to the test items generated by the EAQC, students had to generate two open ended questions, two completion exercises, two single choice questions, and two multiple-choice questions, respectively. Example
concepts and example questions for each of the four question types concerning a different topic were provided. Participants were allowed to use the text while working on this task. This learning activity lasted about 40 minutes. Subsequently, participants again had to attend a test without any help. Contrariwise to the first test, this test included eight prepared questions and lasted 15 minutes. Four questions in this test based on the EAQC; four had been generated by human.

After a further break in the second learning activity, participants were asked to evaluate concepts and questions that had been generated beforehand by the EAQC or by human. In total, 56 concepts and 24 questions (six per each of the four question types) had to be evaluated. From the 56 concepts, 49 had been extracted by the EAQC (highest ranked by the tool) and seven by human. The 49 automatically extracted concepts corresponded to the suitable phrases calculated by the EAQC in descending order from the text (see Section 3 for details). From the 24 questions to be evaluated during the study, 16 questions had been generated by the EAQC and eight questions had been generated by human. The 16 automatically generated test items (four per each question type) based on the four highest ranked concepts that had been extracted by the EAQC.

Participants were asked to evaluate the relevance of a concept using a 5-point Likert scale (1 = not relevant at all; 5 = very relevant). The quality measure for assessing the questions was derived from the observation matrix of Canella, Ciancimino and Campos (2010). This observation matrix originally consisted of the pertinence, level, terminology, and the interdisciplinarity regarding test items created by students. In our context the interdisciplinarity is not appropriate due to the usage of patterns and the focus on specific topics. Therefore we adapted the procedure to evaluate the quality of the automatically or manually generated questions. Participants were asked to evaluate the questions with respect to the following criteria, again using a 5-point Likert scale (1 = very bad; 5 = very good):

- Pertinence: relevancy of a question in the given context
- Level: level of difficulty of a question
- Terminology: appropriateness of the words chosen
- Answer: quality of the reference answer
- Distractors: quality of the listed distractors (for multiple-choice items only)

The order of the concepts and questions to be evaluated was randomized. This second learning activity lasted approximately 45 minutes. At the end of the evaluation task, students had to fill in a questionnaire in which they were asked to answer more general questions about the task (e.g., how difficult it was to generate and evaluate the questions, respectively, or whether the time schedule for each task was appropriate). In total, the whole experiment lasted approximately three hours. Students were also asked to evaluate further questions for homework (results are not included to the analysis presented here).

5.2 Results

In the following, we concentrate on the students’ evaluation of the concepts and the questions in the second learning activity. We first investigated the quality of the concepts extracted by the EAQC by comparing those concepts with manually generated concepts. The mean rating for the concepts extracted by the EAQC was 2.6 (SD = 0.4), for manually extracted concepts it was 4.0 (SD = 0.6; see Figure 7). A two-tailed t-test for dependent measures showed that this difference was reliable, \( t(28) = 14.87; p < .001 \). This means that students evaluated automatically extracted concepts as less relevant compared to concepts extracted by human. However, when we only investigate the relevance of the seven highest ranked concepts provided by the EAQC, mean ratings for the automatically extracted concepts increased to 3.9 (SD = 0.3; Figure 8). In this case, ratings for concepts extracted by the EAQC were equal to concepts extracted by human, \( t(28) = 1.21, p = 0.23 \), meaning that the most suitable automatically extracted concepts were as relevant as manually extracted concepts.

Based on the automatic concept extraction (see Section 3) it can be assumed that the perceived relevance of the concepts provided by the EAQC decreases with their ranking; i.e., we expected that lower ranked concepts after the extraction phase should be evaluated worse compared to the higher ranked concepts. We investigated this assumption by comparing the mean ratings for the first half of the automatically extracted concepts (higher ranked concepts) with the second half of the concepts (lower ranked concepts). A two-tailed t-test for dependent measure showed that students evaluated higher ranked concepts extracted by the EAQC indeed better compared to lower ranked concepts.
Taken together these results showed that the concepts extracted by the EAQC differ as expected in their relevance: Higher ranked concepts were perceived as more relevant compared to lower ranked concepts. Furthermore, these automatically extracted higher ranked concepts did not differ in their relevance from manually extracted concepts.

Figure 7: Mean ratings for concepts extracted by the EAQC compared to manually extracted concepts. Error bars represent the standard error.

Figure 8: Mean ratings for the seven highest ranked concepts extracted by the EAQC and the seven concepts extracted manually. Error bars represent the standard error.

Before discussing the results of the concept analysis more in detail, we present the analysis regarding the quality of the questions provided by the EAQC. For this analysis we only investigated the questions evaluated in the second learning activity because these questions based on the highest ranked
concepts by the EAQC. Hence, we assumed that these questions should be evaluated as relevant as the manually created questions. Table 2 shows examples for “good” and “bad” questions with respect to each question type. We defined a question as “good” regarding an evaluation criterion, when the average rating for this criterion was above 3.5. The respective criterion is presented in parentheses in Table 2. Accordingly, a question was “bad” regarding a specific criterion when the mean rating was below 3.0. For instance, the “good” open-ended question presented in the example received higher ratings regarding its pertinence and terminology; the “bad” multiple-choice question received lower ratings regarding its terminology and its distractors.

**Table 2**: Examples of “good” and “bad” questions by the EAQC for each question type, the respective evaluation criteria which were evaluated best and worst are presented in parentheses. To simplify matters, we did not include answers for open ended questions

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Good Question</th>
<th>Bad Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open ended</td>
<td>What do you know about Modern NLP algorithms in the context of Natural language processing? (Pertinence &amp; Terminology)</td>
<td>What do you know about Natural Language processing in the context of Natural language processing? (Terminology)</td>
</tr>
<tr>
<td>Single choice</td>
<td>Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. [true] (Pertinence &amp; Terminology)</td>
<td>However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in those languages trade [correct: text] edition segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language. (Terminology)</td>
</tr>
<tr>
<td>Completion exercise</td>
<td>[...] Little further research in machine translation was conducted until the late 1980 s, when __________ were developed. [...] Answer: the first statistical machine translation systems (Answer)</td>
<td>__________ (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. [...] Answer: Natural Language processing (Level)</td>
</tr>
<tr>
<td>Multiple choice</td>
<td>[...] Little further research in machine translation was conducted until the late 1980 s, when __________ were developed. [...] A1: the first statistical machine translation systems A2: the first statistical robotics systems A3: the first statistical mt systems (Pertinence)</td>
<td>[...] However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in __________ is a significant task requiring knowledge of the vocabulary and morphology of words in the language. A1: those hyponyms text segmentation A2: those indications text segmentation A3: those languages text segmentation A4: those expressive styles text segmentation (Terminology &amp; Distractors)</td>
</tr>
</tbody>
</table>

Figure 9 shows the comparison between manual and automatically created test items (averaged across question types) regarding the five evaluation criteria (i.e., pertinence, terminology, level, answer, and distractor, respectively) described before. Ratings were generally high with an average of $M = 3.4$ ($SD = 0.4$) for automatically generated questions and $M = 3.7$ ($SD = 0.3$) for manually generated questions. We compared questions created by EAQC and manually created questions for each quality criterion by computing individual two-tailed t-tests for depended measures. Results showed that mean ratings for questions created by EAQC did not differ from the manually created questions regarding pertinence, level, and answer (all $p$’s > .05, Bonferroni corrected). However, regarding terminology and quality of the distractors, questions created by the EAQC were rated worse compared to manually created questions (all $p$’s < .001).

Although comparison between the two conditions (i.e., automatically vs. manually created questions) should be interpreted with caution, because there were less questions created by humans than by the EAQC, results nevertheless suggest, that the quality of the questions created by the EAQC is quite good. As expected from the analysis of the underlying concepts, results indicate that the questions provided by the AGQ were as relevant as questions provided by humans. This is further evidence that the key concepts extracted by the EAQC and hence, the questions that base on these concepts are
indeed equally relevant for the students. However, further experimentation is necessary in order to evaluate the quality of questions that base on less suitable (i.e., lower ranked) concepts.

Figure 9: Comparison of manually and automatically created questions with respect to the defined evaluation criteria. Error bars represent the standard errors.

Furthermore, results also showed that the level of the questions and the provided answers seem to fulfill the needs of the students. Regarding these criteria of the items’ difficulty and the answers, there was no difference between automatically and manually created questions. However, students’ perception of the terminology and the quality of the distractors created by the EAQC was worse compared to their perception of the same aspects regarding manually created questions. A closer look to the data suggests that the terminology was worse especially for completion exercises and multiple choice questions. This is insofar somewhat surprising as the terminology of those question types - when automatically created - did not differ that much from the terminology of the original sentences in the text. For instance, a completion exercise is created by using an existing sentence or paragraph of the text, leaving blank the main concept (= answer) (see also Table 2). Perhaps students are not that familiar with such a style. For instance, when students were asked to create themselves completion exercises and multiple-choice questions during the first learning activity, they typically constructed new sentences and did not simply use the existing ones. Hence, it is possible that not the terminology of the questions per se but their terminology in context of questioning is inappropriate. In any case, further experimentation is necessary to investigate this issue in more detail. For instance, students could be asked to define why the terminology of a question is inappropriate or how it could be improved.

Results also showed that the quality of the distractors provided by the EAQC was worse compared to human created questions. Automatic generation of distractors is still very challenging. Previous research suggests that the chosen distractors should be as semantically close to the correct answer as possible (Mitkov, Ha, & Karamanis, 2005). Our current approach builds on antonyms and related terms on concept or word level. Improvements could be gained by more carefully choosing distractors which we are currently working on. Another alternative for improvements could be the deep study of the process of distractor creation by subject domains in order to implement a similar process chain in the tool. Hence, also in this case further experimentation is necessary in order to create appropriate distractors for multiple choice questions. Furthermore, it might be worth investigating why a specific distractor is suitable or not in order to define enhanced criteria for the improvement of our tool.

Finally, the analysis of the automatically extracted concepts showed that not all 49 concepts extracted by the AGC were equal in their relevance. On the one hand, this is in accordance with the concept...
extraction strategy described in Section 3: The automatically extracted concepts are ranked regarding their suitability and are therefore a priori not expected to be equally relevant at all. On the other hand, however, this nevertheless raises - from a pedagogical viewpoint - three important questions. First, when exactly is a concept relevant or not? Second, what is the appropriate number of concepts that should be extracted in general so that only “relevant” concepts are used for question generation? Third, is it perhaps also worth providing questions that base on “less relevant” concepts? Regarding the first two objections, analysis of one task of the first learning activity of the study showed that students themselves extracted 17.1 concepts on average (SD = 10.3); ranging from 5 to 41 extracted concepts per student. Note at this point that the students were asked to extract the “main” concepts of the text; i.e., to extract such concepts they perceive as relevant. The variance in the number of self-extracted concepts indicates that there are big individual differences between students. Such individual differences should also be taken into account by the EAQC when questions are automatically created. As described before, the user has the possibility to add or deselect phrases during the phase of the automatic concept extraction. In doing so, the EAQC already supports the creation of questions on the basis of the individual students’ requests. A further improvement of the tool could be that a user simply enters relevant concepts (based on his or her individual viewpoint which concepts are relevant) into the system to receive questions from the EAQC. Such an approach would also support the benefit of the EAQC with respect to self-regulated learning activities. However, students sometimes might face the problem that they cannot estimate, which concepts are relevant and which are not. In this case they would miss important concepts for question creation, which, in turn, might impair their learning progress. Therefore, also “less relevant” concepts and the resulting questions might be valuable for a deeper understanding of the learning content. Once again, investigating these issues will be one challenge in future studies.

6. Conclusions and future work

Assessment has to be seen as an integrated and important activity in the learning process. In particular modern educational approaches - such as self-directed or exemplary learning - and personalized learning activities cause a tremendous effort or make it even impossible to prepare appropriate and individualized test items, assess them and provide feedback. To overcome this problem, we advocate an approach which automatically creates test items from learning content, administer knowledge assessment and provide feedback.

We have introduced a concept and prototype implementation, that is capable of handling various text formats and WWW resources, that annotates the corpus using GATE, that applies statistical, semantic and structural methods for identifying key concepts. Based on these concepts the Enhanced Automatic Question Creator (EAQC) generates open ended, single choice, multiple-choice and completion exercises and exports those into QTI items. The evaluation confirmed first promising results and showed the applicability of the system. Encountered problems include (a) the high time complexity for text annotation and WordNet-based operations, (b) problems with specific structures of the content and versions of file formats, (c) partly inappropriate concepts selection due to lack of common sense knowledge and domain knowledge, and (d) the quite low quality of selected distractors.

On the technical level, future work include improvements of better dealing with different content structures, applying common sense and domain knowledge as well as to improve the process of the distractor selection. On the cognitive science and pedagogic level, further pilot studies and evaluations in concrete learning scenarios will be performed.

Acknowledgements

This research is partially supported by the European Commission under the Collaborative Project ALICE "Adaptive Learning via Intuitive/Interactive, Collaborative and Emotional System", VII Framework Program, Theme ICT-2009.4.2 (Technology-Enhanced Learning), Grant Agreement no. 257639. Some of the research work of Klaus Lankmayr and Joachim Weinhofer has been financially supported by an internal research grant of Graz University of Technology. Some contributions of this paper also result of visiting academic activities of Christian Gütl in Australia. The visits were supported and sponsored by School of Information Systems and Curtin Business School, Curtin University, and Institute for Information Systems and Computer Media (IICM) at Graz University of Technology. We are grateful to Isabella Pichlmair and Claudia Pieslinger for their support in the analysis of the study.
References


Methodology for Evaluating Quality and Reusability of Learning Objects

Eugenijus Kurilovas\textsuperscript{1,2}, Virginija Bireniene\textsuperscript{2,3} and Silvija Serikoviene\textsuperscript{1,2}
\textsuperscript{1}Vilnius University Institute of Mathematics and Informatics, Vilnius, Lithuania
\textsuperscript{2}Centre of Information Technologies in Education, Vilnius, Lithuania
\textsuperscript{3}Klaipeda Simon Dach Basic School, Lithuania

\texttt{Eugenijus.Kurilovas@itc.smm.lt}
\texttt{miride@gmail.com}
\texttt{silvija.serikoviene@gmail.com}

Abstract: The aim of the paper is to present the scientific model and several methods for the expert evaluation of quality of learning objects (LOs) paying especial attention to LOs reusability level. The activities of eQ Net Quality Network for a European Learning Resource Exchange (LRE) aimed to improve reusability of LOs of European Schoolnet’s LRE service for schools are analysed in more detail. As a pan-European service, the LRE particularly seeks to identify LOs that can “travel well” (i.e., reusable) across national borders and can be used in a cultural and linguistic context different from the one in which they were created. The primary aim is to improve the quality of LOs in LRE. eQNet is doing this by establishing a network consisting of researchers, policy makers, and practitioners (teachers) that develops and applies “travel well” quality criteria to both existing LRE content as well as that to be selected in future from national repositories. The vision driving the LRE is that a significant percentage of high quality LOs developed in different countries, in different languages and to meet the needs of different curricula can be re-used at European level. The main problem of all existing approaches in the area is a high level of the expert evaluation subjectivity. The authors analyse several scientific approaches, theories, methods and principles to minimise the subjectivity level in expert evaluation of LOs quality, namely: (1) multiple criteria decision analysis approaches for identification of quality criteria, (2) technological quality criteria classification principle, (c) fuzzy group decision making theory to obtain evaluation measures, (d) normalisation requirement for criteria weights, (e) scalarisation method and (f) trapezoidal fuzzy method for LOs quality optimisation. The authors show that the complex application of these approaches could significantly improve the quality of expert evaluation of LOs and noticeably reduce the expert evaluation subjectivity level. The paper also presents several examples of practical application of these approaches for LOs quality evaluation for Physics and Mathematics subjects.

Keywords: learning objects, multiple criteria decision analysis, quality evaluation, reusability, optimisation

1. Introduction: Evaluation of learning objects in eQNet project

The aim of the paper is to present the scientific model and several methods for the expert evaluation of quality of learning objects (LOs) paying especial attention to their reusability level. This is also one of the main objectives of a new eQNet (eQNet 2011) project.

eQNet is a three-year (September 2009-2012) Comenius Multilateral Network funded under the European Commission’s Lifelong Learning Programme. The project is coordinated by European Schoolnet (EUN) and involves 9 Ministries of Education or agencies nominated to act on their behalf. The primary aim is to improve the quality of learning objects (LOs) in European Schoolnet’s Learning Resource Exchange – LRE (LRE 2011) which currently offers almost 130,000 LOs and assets from over 25 providers. As a pan-European service, the LRE particularly seeks to identify LOs that “travel well” (i.e., reusable) across national borders and can be used in a cultural and linguistic context different from the one in which they were created (eQNet 2011).

eQNet is doing this by establishing a network consisting of researchers, policy makers and practitioners (teachers) that develops and applies “travel well” quality criteria to both existing LRE content as well as that to be selected in future from national repositories. The vision driving the LRE is that a significant percentage of high quality LOs developed in different countries, in different languages and to meet the needs of different curricula can be re-used at European level. eQNet provides a forum for joint reflection and co-operation related to the exchange and re-use of educational content and allows network members to:

- Better share information and expertise particularly related to “travel well” quality criteria (i.e., pedagogical, technical and intellectual property rights (IPR) factors);
- Develop new frameworks to improve the quality of LOs and metadata in both national repositories and the LRE, including the growing volume of user-generated content and metadata, as well as to

Reference this paper as: Kurilovas, E., Bireniene, V. and Serikoviene, S.” Methodology for Evaluating Quality and Reusability of Learning Objects” The Electronic Journal of e-Learning Volume 9 Issue 1 2011, (pp39-51), available online at \texttt{www.ejel.org}
improve the multilinguality of LRE content as a result of the translation of metadata, making use, where appropriate, of automatic metadata translation approaches and technologies;

- Enable schools to participate in a Community of Practice related to the use LOs at European level.

Major results will include:

- The development of “travel well” quality criteria to more easily identify LOs with the potential for cross-border use (this work package is coordinated by the Lithuanian partner, in particular by the author of the paper);

- The practical application by teachers of these criteria to >3,500 LOs in the LRE;

- ‘Showcases’ of the best of these LOs in a “travel well” section of the LRE portal;

- Where necessary, the enrichment of selected LOs with new or better metadata;

- A Community of Practice for teachers around these LOs (eQNet 2011).

2. Methodology of the research

One of the main features achieving the high LOs effectiveness and efficiency level is LOs reusability. The need for reusability of LOs has at least three elements (McCormick et al. 2004, Kurilovas 2009):

1. Interoperability: LO is interoperable and can be used in different platforms.

2. Flexibility in terms of pedagogic situations: LO can fit into a variety of pedagogic situations.

3. Modifiability to suit a particular teacher’s or student’s needs: LO can be made more appropriate to a pedagogic situation by modifying it to suit a particular teacher’s or student’s needs.

Reusability of LOs (or their ability to “travel well” between different contexts and education systems) is considered by the authors as a part of the overall quality of LOs. This means that any high quality LO has some reusability level (or potential to “travel well”), but this does not mean that any reusable LO is quality one.

The main problem analysed in the paper is how to establish

1. a ‘proper’ set of LOs “travel well” quality evaluation criteria that should reflect the objective scientific principles of construction a model (criteria tree) for LOs “travel well” quality evaluation, and

2. ‘proper’ methods for evaluation of LOs “travel well” quality.

According to (Oliver 2000), evaluation can be characterised as “the process by which people make judgements about value and worth”. In the context of learning technology this judgement process is complex and often controversial. Although the notion of evaluation is rooted in a relatively simple concept, the process of judging the value of learning technology is complex and challenging. Quality evaluation is defined as “the systematic examination of the extent to which an entity (part, product, service or organisation) is capable of meeting specified requirements” (ISO/IEC 1999). Expert evaluation is referred here as the multiple criteria evaluation of LOs aimed at selection of the best alternatives (i.e., LOs) based on score-ranking results (Kurilovas and Dagiene 2009a).

According to (Zavadskas and Turskis 2010), there is a wide range of multiple criteria decision making (MCDM) problem solution techniques, varying in complexity and possible solutions. Each method has its own strength, weaknesses and possibilities to be applied. But, according to (Zavadskas and Turskis 2010), there are still no rules determining the application of multi-criteria evaluation methods and interpretation of the results obtained.

If the set of decision alternatives (LOs) is assumed to be predefined, fixed and finite, then the decision problem is to choose the optimal alternative or, maybe, to rank them. But usually the experts have to deal with the problem of optimal decision in the multiple criteria situation where the objectives are often conflicting. In this case, an optimal decision is the one that maximises the expert’s utility.
These principles of identification of quality evaluation criteria have been analysed in multiple criteria decision analysis (MCDA) theory related research works (e.g., (Belton and Stewart 2002)).

Evaluation of LOs quality is a typical case where the criteria are conflicting, i.e., LOs could be very qualitative against several criteria, and not qualitative against the other ones, and vice versa. Therefore, the authors propose to use MCDA approach for creation of LOs quality evaluation model.

LOs multiple criteria evaluation method used by the authors in eQNet is referred here as the experts’ additive utility function represented by formula (1) below including LOs evaluation criteria, their ratings (values) and weights (Kurilovas and Serikoviene 2010).

This method is well-known in the theory of optimisation methods and is named “scalarisation method”. A possible decision here could be to transform multi-criteria task into one-criterion task obtained by adding all criteria together with their weights. It is valid from the point of view of the optimisation theory, and a special theorem exists for this case (Kurilovas and Serikoviene 2010).

Therefore, here we have the experts’ additive utility function:

\[ f(X) = \sum_{i=1}^{m} a_i f_i(X), \quad \sum_{i=1}^{m} a_i = 1, \]

where \( f_i(X_j) \) is the rating (i.e., non-fuzzy value) of the criterion \( i \) for the each of the examined LOs alternatives \( X_j \). The weights here should be ‘normalised’ according to the ‘normalisation’ requirement

\[ \sum_{i=1}^{m} a_i = 1, \quad a_i > 1. \]

According to (Zavadskas and Turskis 2010), the normalisation aims at obtaining comparable scales of criteria values. The major is the meaning of the utility function (1) the better LOs meet the quality requirements in comparison with the ideal (i.e., 100%) quality.

3. Literature analysis and research results

This section is aimed to apply the aforementioned scientific approaches in order:

(1) to propose a suitable scientific model for evaluation of quality of LOs,
(2) to propose suitable scientific methods for evaluation of quality of LOs, and
(3) to present the experimental evaluation results using the proposed evaluation model and methods.

3.1 Learning objects quality evaluation model

The following principles of identification of quality evaluation criteria are relevant to all MCDA approaches (Belton and Stewart 2002):

(1) Value relevance: Are the decision makers able to link the concept to their goals, thereby enabling them to specify preferences which relate directly to the concept?

(2) Understandability: It is important that decision makers have a shared understanding of concepts to be used in an analysis.

(3) Measurability: All MCDA implies some degree of measurement of the performance of alternatives against specified criteria, thus it must be possible to specify this in a consistent manner. It is usual to decompose criteria to a level of detail which allows this.

(4) Non-redundancy: Is there more than one criterion measuring the same factor? When eliciting ideas often the same concept may arise under different headings. One can easily check for criteria
which appear to be measuring the same thing by calculating a correlation coefficient if appropriate data is available, or carrying out a process of matching as associated with analysis of repertory grids.

(5) Judgmental independence: Criteria are not judgementally independent if preferences with respect to a single criterion, or trade-offs between two criteria, depend on the level of another.

(6) Balancing completeness and conciseness: A number of authors note that desirable characteristics of a value tree are that it is complete, i.e., that all important aspects of the problem are captured, and also that it is concise, keeping the level of detail to the minimum required.

(7) Operationality: The model is usable with reasonable effort – that the information required does not place excessive demands on the decision makers. The context in which the model is being used is clearly important in judging the usability of a model.

(8) Simplicity versus complexity: The value tree, or criteria set is itself a simple representation, capturing the essence of a problem, which has been extracted from a complex problem description. The modeller should strive for the simplest tree which adequately captures the problem for the decision maker.

LOs quality evaluation model based on these MCDA criteria identification principles is presented in the Fig. 1. This model consists of eight quality criteria, four of them dealing with technological quality, three – with pedagogical quality of LOs, and one – with IPR issues. This model includes three groups of criteria, namely, technological, pedagogical and IPR criteria.

![Figure 1: LOs quality evaluation model (criteria tree)](image-url)
According to the *technological quality criteria classification principle*, we can divide technological quality criteria into 'internal quality' and 'quality in use' criteria of the educational software (e.g., LOs). 'Internal quality' is a descriptive characteristic that describes the quality of software independently from any particular context of its use, while 'quality in use' is evaluative characteristic of software obtained by making a judgment based on criteria that determine the worthiness of software for a particular project (Kurilovas and Dagiene 2009a).

Any LOs quality evaluation model (set of criteria) should provide the experts (decision makers) the clear instrumentality who (i.e., what kind of experts) should analyse what kind of LOs quality criteria in order to select the best LOs suitable for their needs. According to aforementioned technological quality criteria classification principle, 'internal quality' criteria should be mainly the area of interest of the software engineers, and 'quality in use' criteria should be mostly analysed by the programmers and users taking into account the users’ feedback on the usability of software (Kurilovas and Serikoviene 2010).

The authors have applied these two principles in their previous papers (Kurilovas and Dagiene 2009a, 2009b, 2009c, Kurilovas and Serikoviene 2010) on technological evaluation of the learning software, and thus have identified a number of LOs technological quality evaluation criteria presented in the technological part of the LOs quality evaluation model (see Fig. 1).

On the other hand, the authors have analysed a number of existing models (sets of quality evaluation criteria) for evaluation of pedagogical quality of LOs, e.g., (Becta 2007, Leacock and Nesbit 2007, MELT 2008, Vargo et. al. 2003).

Suitable pedagogical reusability criteria based on MCDA principles (Belton and Stewart 2002) are:

1. Interactivity, strong visual structure (animations, images and short videos are travelling best).
2. Language independence or low language dependence (easily translatable) or multilinguality.
3. Ease of use, intuitiveness.

Intellectual property rights (IPR) criterion should also be considered here (Kurilovas and Bireniene 2010).

The authors’ analysis has shown that the model presented in Fig. 1 fits all MCDA criteria identification principles. Taking into account Non-redundancy, Judgmental independence, Balancing completeness and conciseness, Operationality, and Simplicity versus complexity MCDA criteria identification principles, the authors consider that the following eight LOs evaluation criteria should construct the comprehensive LOs quality criteria tree (see Fig. 1):

**Technological quality criteria:**

'Internal quality' criteria:

1. Technological reusability:
   - Interoperability: Metadata accuracy; Compliance with the main import/export standards (e.g., IMS CP, IMS CC, SCORM 2004).
   - Decontextualisation: Is LO indivisible (atomic)? – LO aggregation (granularity) level; Is LO modular (i.e., are the parts of a content item fully functional on their own?).
   - Cultural and learning diversity (Adaptability): LO flexibility (LO can be modified, for instance from a configuration file, from a plain text file or because it is provided along with its source code or an authoring tool); LO internationalisation level; LO suitability for localisation.
   - Accessibility (design of controls and presentation formats to accommodate disabled and mobile learners): Is LO designed for all?; Compliance with accessibility standards (W3C) (Kurilovas and Dagiene 2009c).
2. Architecture: Is LO architecture layered in order to separate data, presentation and application logics? (Kurilovas and Dagiene 2009c).
(3) Robustness, technical stability:
- Having help functions that identify common user problems and their solutions.
- Having navigational actions that can be undone.
- Giving quick, visible and audible responses to user actions.
- Allowing the user to exit at any point.
- Not being adversely affected by user experimentation and error. If users do experience an error they should be able to recover quickly and, where appropriate, be informed about the nature of the error (Becta, 2007)

‘Quality in use’ criterion:

(4) Design and usability (design of visual and auditory information for enhanced learning and efficient mental processing): Aesthetics; Navigation; User-friendly interface; Information structuring; Personalisation (Kurilovas and Dagiene 2009c).

These LOs technological quality criteria are included into a majority of the aforementioned LOs evaluation models. ‘Interoperability’ and ‘Accessibility’ criteria being independent criteria in e.g., (Leacock and Nesbit 2007) or (Becta 2007) are included as sub-criteria into ‘Technical reusability’ criterion in the presented model. There are several reasons for this, e.g., both ‘Interoperability’ and ‘Accessibility’ criteria deal with international interoperability standards and specifications, both influence LO technical reusability level in different repositories and platforms, etc. MCDA Non-redundancy principle is applied here.

Pedagogical quality criteria:
(5) Interactivity, strong visual element: e.g., LOs include animations, images, short videos and simulations that are self-explanatory or have just a few text labels or icons/buttons for start, stop, etc.; strong visual structure (MELT 2008).

(6) Language independence: LO is not text-heavy; LOs may have little or no text; or low language dependence (easily translatable); or LOs are multilingual, i.e., LOs have been designed to be language customisable and are already offered in more than one language (MELT 2008).

(7) Ease of use, intuitiveness:
- Users can find their way through the resource almost intuitively; they can broadly understand what is the intended learning objective or topic (MELT 2008).
- LOs provide appropriate guidance, where necessary, for learners and/or practitioners.
- LOs make appropriate assumptions about the ICT skills of users, both learners and practitioners, or provide straightforward guidance on this.
- LOs not present a barrier or impede the learning experience (Becta 2007).

These LOs pedagogical quality criteria are included into the analysed LOs evaluation models (Leacock and Nesbit 2007, Becta 2007, MELT 2008, Vargo et al. 2003).

There are also several criteria included into the aforementioned models. They are: ‘Content quality’ (Leacock and Nesbit 2007), ‘Match to the curriculum’, ‘Assessment’, ‘Learner engagement’, and ‘Innovative approaches’ (Becta 2007).

In the authors’ opinion, ‘Content quality’ criterion should not be included into the proposed LOs quality evaluation model. The main reason for this is “a need for a common (more narrow) definition of what is, and what is not a LO” (Paulsson and Naeve 2006). Therefore, since a LO is “any digital resource that can be reused to support learning” (Wiley 2000), we can conclude that scientific content of any LO should be relevant, accurate and trustworthy, in other case the digital resource could not be used and reused “to support learning”.

‘Match to the curriculum’ criterion could be suitable for nationally recognised quality criteria of learning resources, but taking into account the “more narrow definition” (Paulsson and Naeve 2006) of LOs, we should consider only the resources “that can be reused” to support learning (Wiley 2000). Since
there are different curricula in different countries, reusable resources should be curriculum independent. Reusable LOs could be often used in other pedagogic situations and learning scenarios that have been planned by LOs authors. Such approach has been, e.g., applied by the author’s leaded team in FP6 CALIBRATE project (CALIBRATE 2008) where Lithuanian teachers have been created and implemented their own lesson plans using foreign LOs from LRE.

Criteria such as ‘Assessment to support learning’ and ‘Robust summative assessment’ (Becta 2007) often do not fit the LO reusability principle – they are mostly suitable for entire learning courses with high semantic density and aggregation level.

Criteria such as ‘Learner engagement’ and ‘Innovative approaches’ are interconnected on the one hand, and are closely related to ‘Interactivity’ and ‘Intuitiveness’ criteria included into the proposed model, on the other.

**IPR quality criterion:**

(8) Open license, free to use, open code: Licensing (clear rules, e.g., compliance with Creative Commons); Economic efficiency – Cost versus Quality taking into account probable LO reusability level (Kurilovas 2007).

### 3.2 Learning objects quality evaluation methods

The widely used measurement criteria of the decision attributes’ quality are mainly qualitative and subjective. Decisions in this context are often expressed in natural language, and evaluators are unable to assign exact numerical values to the different criteria. Assessment can be often performed by linguistic variables: ‘bad’, ‘poor’, ‘fair’, ‘good’ and ‘excellent’. These values are imprecise and uncertain: they are commonly called ‘fuzzy values’. Integrating these different judgments to obtain a final evaluation is not evident (Kurilovas and Serikovienė 2010). Therefore, the authors have proposed to use *fuzzy group decision making theory* (Ounaies et al. 2009) to obtain final assessment measures. The fuzzy numbers are: (1) triangular fuzzy numbers, (2) trapezoidal fuzzy numbers, and (3) bell-shaped fuzzy numbers. In the presented paper, the authors use *triangular and trapezoidal fuzzy numbers* for evaluating quality and reusability of LOs.

**Use of triangular fuzzy numbers**

According to (Zhang Li Li and Cheng De Yong 1992), triangular fuzzy numbers (TFNs) are a class of the fuzzy set representation. A triangular fuzzy number is expressed by three real numbers $M = (l, m, u)$; the parameters $l$, $m$ and $u$, respectively, indicate the lower, the mean and the upper possible values. TFNs membership functions are as follows:

$$
\mu_M(x) = \begin{cases} 
\frac{x - l}{m - l}, & \text{if } x \in [l, m], \\
\frac{x - m}{m - u}, & \text{if } x \in [m, u], \\
0, & \text{if } x \not\in [l, u]. 
\end{cases}
$$

Figure 2 illustrates triangular fuzzy numbers.

![Figure 2: Triangular fuzzy numbers](image)
Conversion of these qualitative values into fuzzy numbers is shown in Table 1.

Table 1: Linguistic variables conversion into triangular fuzzy numbers

<table>
<thead>
<tr>
<th>Linguistic variables</th>
<th>Triangular fuzzy numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>(0.700, 0.850, 1.000)</td>
</tr>
<tr>
<td>Good</td>
<td>(0.525, 0.675, 0.825)</td>
</tr>
<tr>
<td>Fair</td>
<td>(0.350, 0.500, 0.650)</td>
</tr>
<tr>
<td>Poor</td>
<td>(0.175, 0.325, 0.475)</td>
</tr>
<tr>
<td>Bad</td>
<td>(0.000, 0.150, 0.300)</td>
</tr>
</tbody>
</table>

Therefore, in the case of using average triangular fuzzy numbers, linguistic variables conversion into non-fuzzy values of the evaluation criteria should be as follows: 'excellent'=0.850; 'good'=0.675; 'fair'=0.500; 'poor'=0.325; 'bad'=0.150 (Kurilovas and Serikoviene 2010).

The weight of the evaluation criterion reflects the experts’ opinion on the criterion’s importance level in comparison with the other criteria for the particular needs. For example, for the most simple (general) case, when all LOs evaluation criteria are of equal importance (i.e., we pay no especial attention to LOs reusability criteria), the experts could consider the equal weights a_i = 0.125 according to the normalisation requirement (2).

But if we pay especial attention to LOs reusability criteria, we can, e.g., consider the increased weights for the 1st and 6th LOs quality evaluation criteria (see Fig. 1 and Tables 2, 3 below), because these criteria deal with LOs reusability mostly. In this case, the authors while implementing eQNet have decided to apply increased weights that are twice higher in comparison with the other ones (i.e., 0.2), and all other criteria weights according to normalisation requirement (2) should be equal 0.1.

Lithuanian Physics expert teacher (the co-author of the paper) has applied the presented evaluation model and triangular fuzzy numbers method in eQNet project (see Table 2 below).

A number of probably qualitative reusable Physics LOs have been identified in Lithuanian LOs repositories and evaluated against the model and method presented above (see formula (1)).

There are several examples of these LOs presented in Table 2:

(1) LO_1: Light diffraction (available online at http://mkp.emokykla.lt/imo/lt/mo/330/);
(2) LO_2: Photo effect (available online at http://mkp.emokykla.lt/imo/lt/mo/367/); and
(3) LO_3: Isobar process (available online at http://mkp.emokykla.lt/imo/lt/mo/395/).

Table 2: Results of experimental evaluation of Physics LOs general quality (q) and “travel well” quality (twq) using triangular fuzzy numbers method

<table>
<thead>
<tr>
<th>LOs evaluation criteria</th>
<th>LO_1q</th>
<th>LO_2q</th>
<th>LO_3q</th>
<th>LO_1twq</th>
<th>LO_2twq</th>
<th>LO_3twq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technological criteria:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Technological reusability</td>
<td>0.850</td>
<td>0.850</td>
<td>0.850</td>
<td>0.1700</td>
<td>0.1700</td>
<td>0.1700</td>
</tr>
<tr>
<td>2. Design and usability</td>
<td>0.850</td>
<td>0.850</td>
<td>0.850</td>
<td>0.0850</td>
<td>0.0850</td>
<td>0.0850</td>
</tr>
<tr>
<td>3. Working stability</td>
<td>0.850</td>
<td>0.850</td>
<td>0.850</td>
<td>0.0850</td>
<td>0.0850</td>
<td>0.0850</td>
</tr>
<tr>
<td>4. Architecture</td>
<td>0.675</td>
<td>0.675</td>
<td>0.675</td>
<td>0.0675</td>
<td>0.0675</td>
<td>0.0675</td>
</tr>
<tr>
<td>Pedagogical criteria:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Interactivity level</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.0500</td>
<td>0.0500</td>
<td>0.0500</td>
</tr>
<tr>
<td>6. Language independence</td>
<td>0.850</td>
<td>0.675</td>
<td>0.500</td>
<td>0.1700</td>
<td>0.1350</td>
<td>0.1000</td>
</tr>
<tr>
<td>7. Ease of use, intuitiveness</td>
<td>0.850</td>
<td>0.850</td>
<td>0.850</td>
<td>0.0850</td>
<td>0.0850</td>
<td>0.0850</td>
</tr>
<tr>
<td>IPR criteria:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Open licence, cost</td>
<td>0.850</td>
<td>0.850</td>
<td>0.850</td>
<td>0.0850</td>
<td>0.0850</td>
<td>0.0850</td>
</tr>
<tr>
<td><strong>Evaluation results:</strong></td>
<td><strong>0.7844</strong></td>
<td><strong>0.7625</strong></td>
<td><strong>0.7406</strong></td>
<td><strong>0.7975</strong></td>
<td><strong>0.7625</strong></td>
<td><strong>0.7275</strong></td>
</tr>
</tbody>
</table>
These results mean that LO₁ meets 78.44% quality (q) in comparison with the ideal, LO₂ – 76.25%, and LO₃ – 74.06%. They also mean that LO₁ meets 79.75% “travel well” quality (twq) in comparison with the ideal, LO₂ – 76.25%, and LO₃ – 72.75%.

Therefore, using triangular fuzzy numbers method, one could see that LO₁ is the best alternative (among the evaluated) both from general quality and “travel well” quality points of view.

Lithuanian Mathematics expert teacher (the co-author of the paper) has also applied the presented evaluation model and triangular fuzzy numbers method in eQNet project (see Table 3).

A number of probably qualitative reusable LOs have been identified in Lithuanian LOs repositories and evaluated against the aforementioned model and method (see formula (1)).

There are three examples of these LOs presented in Table 3:
- LO₁: “Coordinate Method” (available online at http://mkp.emokykla.lt/imo/lt/mo/250/);
- LO₂: “Polygon area” (available online at http://mkp.emokykla.lt/imo/lt/mo/431/); and
- LO₃: “Interval Method” (available online at http://mkp.emokykla.lt/imo/lt/mo/316/).

Table 3: Results of experimental evaluation of Mathematics LOs general quality (q) and “travel well” quality (twq) using triangular fuzzy numbers method

<table>
<thead>
<tr>
<th>LOs evaluation criteria</th>
<th>LO₁q</th>
<th>LO₂q</th>
<th>LO₃q</th>
<th>LO₁twq</th>
<th>LO₂twq</th>
<th>LO₃twq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technological criteria:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Technological reusability</td>
<td>0.675</td>
<td>0.850</td>
<td>0.675</td>
<td>0.1350</td>
<td>0.1700</td>
<td>0.1350</td>
</tr>
<tr>
<td>2. Design and usability</td>
<td>0.675</td>
<td>0.850</td>
<td>0.850</td>
<td>0.0675</td>
<td>0.0850</td>
<td>0.0850</td>
</tr>
<tr>
<td>3. Working stability</td>
<td>0.675</td>
<td>0.500</td>
<td>0.675</td>
<td>0.0675</td>
<td>0.0500</td>
<td>0.0675</td>
</tr>
<tr>
<td>4. Architecture</td>
<td>0.675</td>
<td>0.500</td>
<td>0.500</td>
<td>0.0675</td>
<td>0.0500</td>
<td>0.0500</td>
</tr>
<tr>
<td>Pedagogical criteria:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Interactivity level</td>
<td>0.850</td>
<td>0.500</td>
<td>0.325</td>
<td>0.0850</td>
<td>0.0500</td>
<td>0.0325</td>
</tr>
<tr>
<td>6. Language independence</td>
<td>0.675</td>
<td>0.850</td>
<td>0.325</td>
<td>0.1350</td>
<td>0.1700</td>
<td>0.0650</td>
</tr>
<tr>
<td>7. Ease of use, intuitiveness</td>
<td>0.850</td>
<td>0.850</td>
<td>0.500</td>
<td>0.0850</td>
<td>0.0850</td>
<td>0.0500</td>
</tr>
<tr>
<td>IPR criteria:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Open licence, cost</td>
<td>0.850</td>
<td>0.850</td>
<td>0.850</td>
<td>0.0850</td>
<td>0.0850</td>
<td>0.0850</td>
</tr>
<tr>
<td>Evaluation results:</td>
<td>0.7406</td>
<td>0.7188</td>
<td>0.5875</td>
<td>0.7275</td>
<td>0.7450</td>
<td>0.5700</td>
</tr>
</tbody>
</table>

These results mean that LO₁ meets 74.06% general quality (q) in comparison with the ideal, LO₂ – 71.88%, and LO₃ – 58.75%. They also mean that LO₁ meets 72.75% “travel well” quality (twq) in comparison with the ideal, LO₂ – 74.50%, and LO₃ – 57.00%.

Therefore, using triangular fuzzy numbers method, one could see that LO₁ is the best alternative (among the evaluated) from general quality point of view, but LO₂ is the best from “travel well” quality point of view.

These two examples show that the application of the presented model and triangular fuzzy numbers method for evaluation of LOs quality can show both
- Similar results for ‘general quality’ and ‘travel well quality’ of LOs (see Table 2), and
- Different results for ‘general quality’ and ‘travel well quality’ (see Table 3) of different LOs.

Use of trapezoidal fuzzy numbers

A trapezoidal fuzzy number is a fuzzy number represented by four points as follows: \( M = (a, b, c, d) \).

In this case, a membership function can be attached to the level fuzzy function:
same approach to the weights of criteria are presented in Tables 5 and 6 below.

\[
\mu, (x) = \begin{cases} 
0, & \text{if } x < a, \\
\frac{x-a}{b-a}, & \text{if } a \leq x \leq b, \\
1, & \text{if } b \leq x \leq c, \\
\frac{d-x}{d-c}, & \text{if } c \leq x \leq d, \\
0, & \text{if } x > d.
\end{cases}
\]

Figure 3 illustrates trapezoidal fuzzy numbers.

**Figure 3: Trapezoidal fuzzy numbers**

Conversion of these qualitative values into fuzzy numbers is shown in Table 4.

**Table 4: Linguistic variables conversion into trapezoidal fuzzy numbers**

<table>
<thead>
<tr>
<th>Linguistic variables</th>
<th>Trapezoidal fuzzy numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>(0.800, 1.000, 1.000, 1.000)</td>
</tr>
<tr>
<td>Good</td>
<td>(0.600, 0.800, 0.800, 1.000)</td>
</tr>
<tr>
<td>Fair</td>
<td>(0.300, 0.500, 0.500, 0.700)</td>
</tr>
<tr>
<td>Poor</td>
<td>(0.000, 0.200, 0.200, 0.400)</td>
</tr>
<tr>
<td>Bad</td>
<td>(0.000, 0.000, 0.000, 0.200)</td>
</tr>
</tbody>
</table>

Therefore, in the case of using secondary trapezoidal fuzzy numbers, linguistic variables conversion into non-fuzzy values of the evaluation criteria should be as follows: ‘excellent’=1.000; ‘good’=0.800; ‘fair’=0.500; ‘poor’=0.200; ‘bad’=0.000. The results of evaluation of quality of the same Physics and Mathematics LOs identified in Lithuanian LOs repositories using trapezoidal fuzzy numbers and the same approach to the weights of criteria are presented in Tables 5 and 6 below.

**Table 5: Results of experimental evaluation of Physics LOs general quality (q) and “travel well” quality (twq) using trapezoidal fuzzy numbers**

<table>
<thead>
<tr>
<th>LOs evaluation criteria</th>
<th>LO1,q</th>
<th>LO2,q</th>
<th>LO3,q</th>
<th>LO1,twq</th>
<th>LO2,twq</th>
<th>LO3,twq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technological criteria:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Technological reusability</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.2000</td>
<td>0.2000</td>
<td>0.2000</td>
</tr>
<tr>
<td>2. Design and usability</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.1000</td>
<td>0.1000</td>
<td>0.1000</td>
</tr>
<tr>
<td>3. Working stability</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.1000</td>
<td>0.1000</td>
<td>0.1000</td>
</tr>
<tr>
<td>4. Architecture</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.0800</td>
<td>0.0800</td>
<td>0.0800</td>
</tr>
<tr>
<td>Pedagogical criteria:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Interactivity level</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.0500</td>
<td>0.0500</td>
<td>0.0500</td>
</tr>
<tr>
<td>6. Language independence</td>
<td>1.000</td>
<td>0.800</td>
<td>0.500</td>
<td>0.2000</td>
<td>0.1600</td>
<td>0.1000</td>
</tr>
<tr>
<td>7. Ease of use, intuitiveness</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.1000</td>
<td>0.1000</td>
<td>0.1000</td>
</tr>
<tr>
<td>IPR criteria:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Open licence, cost</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.1000</td>
<td>0.1000</td>
<td>0.1000</td>
</tr>
</tbody>
</table>

**Evaluation results:** 0.9125 0.8875 0.8500 0.9300 0.8900 0.8300
These results mean that LO_1 meets 91.25% quality (q) in comparison with the ideal, LO_2 – 88.75%, and LO_3 – 85.00%. They also mean that LO_1 meets 93.00% “travel well” quality (twq) in comparison with the ideal, LO_2 – 89.00%, and LO_3 – 83.00%.

Therefore, using trapezoidal fuzzy numbers method, one could see that LO_1 is the best alternative (among the evaluated) both from general quality and “travel well” quality points of view. The results are similar in comparison with the ones obtained using triangular fuzzy numbers.

**Table 6**: Results of experimental evaluation of Mathematics LOs general quality (q) and “travel well” quality (twq) using trapezoidal fuzzy numbers

<table>
<thead>
<tr>
<th>LOs evaluation criteria</th>
<th>LO_1q</th>
<th>LO_2q</th>
<th>LO_3q</th>
<th>LO_1twq</th>
<th>LO_2twq</th>
<th>LO_3twq</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technological criteria:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Technological reusability</td>
<td>0.800</td>
<td>1.000</td>
<td>0.800</td>
<td>0.1600</td>
<td>0.2000</td>
<td>0.1600</td>
</tr>
<tr>
<td>2. Design and usability</td>
<td>0.800</td>
<td>1.000</td>
<td>1.000</td>
<td>0.0800</td>
<td>0.1000</td>
<td>0.1000</td>
</tr>
<tr>
<td>3. Working stability</td>
<td>0.800</td>
<td>0.500</td>
<td>0.800</td>
<td>0.0800</td>
<td>0.0500</td>
<td>0.0800</td>
</tr>
<tr>
<td>4. Architecture</td>
<td>0.800</td>
<td>0.500</td>
<td>0.500</td>
<td>0.0800</td>
<td>0.0500</td>
<td>0.0500</td>
</tr>
<tr>
<td><strong>Pedagogical criteria:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Interactivity level</td>
<td>1.000</td>
<td>0.500</td>
<td>0.200</td>
<td>0.1000</td>
<td>0.0500</td>
<td>0.0200</td>
</tr>
<tr>
<td>6. Language independence</td>
<td>0.800</td>
<td>1.000</td>
<td>0.200</td>
<td>0.1600</td>
<td>0.2000</td>
<td>0.0400</td>
</tr>
<tr>
<td>7. Ease of use, intuitiveness</td>
<td>1.000</td>
<td>1.000</td>
<td>0.500</td>
<td>0.1000</td>
<td>0.1000</td>
<td>0.0500</td>
</tr>
<tr>
<td><strong>IPR criteria:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Open licence, cost</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.1000</td>
<td>0.1000</td>
<td>0.1000</td>
</tr>
<tr>
<td><strong>Evaluation results:</strong></td>
<td><strong>0.8750</strong></td>
<td><strong>0.8125</strong></td>
<td><strong>0.6250</strong></td>
<td><strong>0.8600</strong></td>
<td><strong>0.8500</strong></td>
<td><strong>0.6000</strong></td>
</tr>
</tbody>
</table>

These results mean that LO_1 meets 87.50% general quality (q) in comparison with the ideal, LO_2 – 81.25%, and LO_3 – 62.50%. They also mean that LO_1 meets 86.00% “travel well” quality (twq) in comparison with the ideal, LO_2 – 85.00%, and LO_3 – 60.00%.

Therefore, using trapezoidal fuzzy numbers method, one could see that LO_1 is the best alternative (among the evaluated) both from general quality and “travel well” quality points of view. The general quality evaluation results are similar, but “travel well” quality evaluation results are a little bit different in comparison with the ones obtained using triangular fuzzy numbers.

In real life situations a teacher is the only suitable expert to decide on quality of LOs, and, therefore, on purposefulness to use these LOs in his / her teaching process in particular school.

In the analysed cases it is clear that Physics teacher should choose LO_1 as the best alternative both from general quality and “travel well” quality points of view.

Mathematics teacher should choose LO_1 as the best alternative from general quality point of view, and either LO_1 or LO_2 – from “travel well” quality point of view since the difference between “travel well” quality evaluation values of LO_1 and LO_2 is very small.

4. Conclusion and recommendations

The research results presented in the paper show that the complex application of the principles of multiple criteria decision analysis for identification of quality evaluation criteria, technological quality criteria classification principle, fuzzy group decision making theory to obtain final evaluation measures, and normalisation requirement for the weights of evaluation criteria, as well as triangular and trapezoidal fuzzy numbers methods for LOs quality optimisation are (1) applicable in real life situations when schools have to decide on purchase of LOs for their education needs, and (2) could significantly improve the quality of expert evaluation of LOs by noticeably reduce of the expert evaluation subjectivity level.

The experimental evaluation results show that the proposed scientific approaches are quite objective, exact and simply to use for selecting the qualitative LOs alternatives in the market.
On the other hand, the proposed LOs “travel well” quality evaluation approach is applicable for the aims of eQNet project in order to select “travel well” LOs from LRE or elsewhere to use them in the other education contexts and countries.

Therefore, these approaches have been recommended by the authors to be widely used by European policy makers, publishers, practitioners, and experts-evaluators both inside and outside eQNet project to evaluate quality and reusability level of LOs.

5. Appendix

The work presented in this paper is partially supported by the European Commission under the LifeLong Learning programme – as part of the eQNet project (project number 502857-LLP-1-2009-1-BECOMENIUS-CNW). The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of data appearing therein.

References


Fluidity in the Networked Society - Self-initiated learning as a Digital Literacy Competence

Karin Tweddell Levinsen
Aarhus University, Copenhagen, Denmark
kale@dpu.dk

Abstract: In the globalized economies e-permeation has become a basic condition in our everyday lives. ICT can no longer be understood solely as artefacts and tools and computer-related literacy are no longer restricted to the ability to operate digital tools for specific purposes. The network society, and therefore also eLearning are characterized by fluidity and the key competence for social actors in this ever changing e-permeated environment is the ability to cope with change - or Castells’ conceptualisation self-programming. Castells’ theory has influenced international definitions of future key competencies. Both lifelong learning and digital literacy understood as "bildung" have emerged as central for the definitions of and standards for future key competencies. However, definitions and standards only tell us about the desired destination and outcome of digital competence building. They tell us nothing about how we may get there. In the educational system ICT and e-learning are becoming an everyday condition and the basic challenge for the educational system is twofold: 1) The actually making of digital literate and self-programming social actors – students and teachers; and 2) How to develop adequate designs for teaching and learning for that purpose. We need research that aims to describe the phenomenology of acquiring digital literacy and self-programming in order to be able to identify relevant learning objectives and scaffolding. Findings from such studies are expected to be relevant for eLearning scenarios as well as for ICT and designs for learning in general. This paper presents a case study that aimed to explore the phenomenological appearance of self-programming as agency and learning among postgraduate students who participated in a specially designed eLearning workshop in the autumn 2009. The findings relate to both the individual and collaborative barriers and proactive strategies that come into play among the students. Drawing on the findings, it is argued that the presented workshop design contributes to the networked society’s design for ICT, teaching and learning, as the design – at least for this small group of students – have proved to support the development of digital self-programming as a sustainable competence. In the autumn 2010 the study will be expanded to a larger group of students.

Keywords: self-programming, lifelong learning, networked society, design for teaching and learning, eLearning

1. Introduction

E-permeation has brought the majority of populations in the globalized economies into an everyday life where ICT is intertwined with almost anything we do in relation to our job, education, public services and society, our friends and family. ICT has become more than artefacts and tools, just as computer-related competencies have become more that the ability to operate digital tools for specific purposes. The structure and organisation of the network society are characterized by fluidity and the basic demand for citizens in this ever changing environment is the ability to cope with change. Consequently, both lifelong learning and digital literacy understood as general education or “bildung” have emerged as central for the definition of future key competencies of the networked society (Tyner 1998, Jewett & Kress 2003, Martin 2006, Katz 2007, Bawden 2008, Levinsen 2009).

Castells (2000) divides employees of the global economy into two dominant types; self-programmable and generic labour. Self-programmable labour is equipped with competencies for lifelong learning, the ability to retrain and to adapt to new conditions and challenges. By contrast, generic labour is both interchangeable and disposable. Castells argue that for a society to remain competitive in the global economy, the educational systems should devote particular efforts to the education of self-programmable individuals. Castells’ theory has had major impact on the international definitions of future key competencies used as guidelines for governmental decisions about education (OECD 2001, European Commission 2003, Rychen & Salganik 2003, Elearning Europa 2005, G8 2006). The job marked adapts faster than societal institutions and job ads related to network- and knowledge production already display that employers’ demand self-programmable labour.

However, both the international definitions and the employers demands are product oriented. Descriptions of what it actually means to act as a lifelong learner or a self-programmable individual are rare and questions that seem obvious are never asked, e.g.:

- What do self-programmable individuals do when they self-program?
- How do adults become self-programmable, if they are not?
How do we ensure that new generations grow up to be self-programmable rather than generic?

It seems as if the formation of key competencies is taken for granted and that the so called digital natives are bound to grow up as self-programmable individuals just because their whole life has been e-permeated. During the last decades, a substantial body of constructivist and social constructivist research in the field of teaching and learning has demonstrated that competencies in dealing with complexity are neither congenital nor readymade for use (Sørensen 1999, Sørensen, Olesen & Audon 2001, Buckingham 2003, Malyn-Smith 2004, Breivik 2005, Levinsen 2006, Levinsen 2009). Additionally, research has documented that ICT and learning do not constitute a simple linear determinist or behavioural relation. Nevertheless, technological determinism persists as the making of self-programmable individuals for the network society seems to rest on the unspoken assumption that ICT is a natural driver for the learning process.

As Castells’ concept self-programming bears connotations to the cognitive psychology’s alignment between the human mind and the computer, I will in the following use the terms self-initiated learners when referring to individuals, and self-initiated learning when referring to the learning outcome in order to underline the constructivist dimension of Castells’ concept. The term self-programming refers in the following reserved to the individuals’ actual performance of self-initiated learning.

In Denmark the consequences of almost two decades of strong focus on the objectives of ICT-implementation rather than on the process of continuous and sustainable implementation and adaptation of ICT have recently become documented. Despite the huge investments in implementation of ICT in primary schools, the message is that ICT is not used in relation to the curriculum and a majority of teachers do not integrate ICT in their every day teaching (Levinsen & Sørensen 2008, EVA 2009). In 2009 the Danish government released New Shared Goals (UVM 2009) for primary schools where part of the objectives and measures depend on active integration of ICT. Both Levinsen & Sørensen and the EVA-report pinpoint implementation-processes and strategies as pivotal for a successful integration of ICT and both reports identify and describe aspects of the school management’s responsibility for nurturing a knowledge sharing organisational culture. Additionally, both reports argue that in-service training has to take on new forms as expressed in this excerpt from the EVA report:

… still necessary to devote particular efforts on the teachers competence building and dedicated support in the everyday practice. The teachers ask for computer literacy courses but they are also aware that former courses did not qualify their teaching practice. The research identifies a need for new models for competence building that aims to integrate ICT in the curriculum subjects and focus on actual use with a contextualised outset in the teachers’ concrete needs (Authors translation from EVA 2009, p. 8)

To some degree we can draw on research into digital natives’ informal approaches to ICT and learning outside school. However, this research only tells us the characteristics of already self-initiated learners and how they do self-programming. According to the literature self-initiated learners who encounter something new, wonder and ask questions. They experiment and explore to figure out about the unknown. They are open, receptive to input and find it natural to share knowledge and to network. They see and transfer potentials between contexts and they are creative and possess a wide repertoire of strategies to explore the unknown. They possess a strong inner drive and motivation to conquer challenges (Sørensen 1999, Oblinger 2003, Malyn-Smith 2004, Dede 2005, Oblinger & Oblinger 2005, Levinsen 2006, Levinsen 2009). The descriptions tell us about the learning objectives or the outcome of self-initiated learning in terms of competence building, but they tell us nothing about how we actually get there. That is, how we actually perform self-programming.

In conclusion, the basic challenge for the society and the educational system is neither the definitions nor the demands for self-initiated or lifelong learners, but the actually making of self-programmable individuals. We need to know how to develop designs for teaching and learning that supports self-programming as competence building. In other words, we need research into the phenomenology of learning self-programming in order to describe aspects of the objectives and designs for teaching and learning that may support self-programming as competence building.

2. Presenting the case

The paper aims to explore the phenomenological appearance of self-programming among postgraduate students who participate in the course Technology-related Workshop at the Danish
postgraduate programme *Designs for ICT, teaching and Learning*. The course took place over 3 months in the autumn 2009 and the activities were based on group work. The course was subdivided into three separate workshops and the study took place during workshop I, session one (Table 1).

### Table 1: Schedule for workshop 1 - technologies

<table>
<thead>
<tr>
<th>Session</th>
<th>Content</th>
<th>Teacher present</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Introduction to the module Workshop 1 - Technologies</td>
<td>The author</td>
</tr>
<tr>
<td>2.</td>
<td>Workshop 1 - Technologies</td>
<td>The author</td>
</tr>
<tr>
<td></td>
<td>Workshop 1 - Technologies Workshop 1 assignment</td>
<td>Self study period</td>
</tr>
<tr>
<td>3.</td>
<td>Workshop 1 - Technologies Presentation of projects and discussion</td>
<td>The author</td>
</tr>
</tbody>
</table>

Three students who considered themselves as flexible and experienced lifelong learners with a high level of digital literacy participated in the study:

- **Amanda**, age 33, primary school teacher – teach visual culture and arts. Additionally, Amanda teaches primary school teachers and adult teachers. She is the author of textbooks in designs for teaching in visual culture and arts.
- **Brenda**, age 45, special training teacher – works with adult dyslexics and their use of digital support in their study, job and everyday lives.
- **Christian**, age 26, primary school teacher – teaches ICT and learning practice at the teacher’ college and works as technical support at the college.

During a full day in workshop I, the students were presented to an entirely new and therefore unknown digital prototype named *Topobo* from MIT media Lab (Raffle, Parkes, & Ishii 2004). They were just given the cardboard box containing the prototype and asked to explore the content and figure out how it worked and how it may be used for learning purposes. The students were instructed to produce individual written logs in their e-portfolios regarding their strategies with special attention to questions like:

- How do I act in order to explore?
- When do I prefer to collaborate or explore on my own?
- Do I experience obstacles or challenges?
- What obstacles/challenges do I encounter and how do I act on them?
- What helps me to bypass obstacles/ solve challenges?
- Do I change my strategy during the day?

Further the students were asked to write down their immediate reflections at the end of the day and reflect on the prototypes learning potentials. For assessment of the course, the students used their e-portfolio and wrote 3-4 pages reflection papers and presented their individual experiences, reflections and learning on the last day of workshop I. For research purpose, the session was video recorded with two cameras: One stationary camera covering the total and a handheld camera for close ups. The recordings were made accessible for the students as empirical data for their reflection paper. The students written material constitute the research data together with the video recordings.

### 2.1 Analytical tools – making learning visible

The students were introduced to a theoretical toolbox with a triple purpose: 1) to make learning visible and shareable among the participants; 2) to scaffold the students’ 2. order reflections on their performance with special attention to their strategies and attitudes when introduced to something unknown; 3) From a phenomenological research perspective the toolbox facilitates the researcher’s insight into the students first-person perspective. For these purposes Castells’ theory of the networked society (2000), Rogers’ Theory of Diffusion of Innovation (1985) and Dreyfus & Dreyfus’ Model of Skill Acquisition (1988) were chosen.
2.1.1 Informal vs formal approach and the adoption of novelties

Castells distinction between self-programmable and generic individuals, described previously, offers a tool to reflect on whether encountering something unknown generates a personal initiative or a need for teaching and external support.

2.1.2 Inner motivation and the adoption of novelties

Rogers’ widely recognized Diffusion of Innovation Theory deals with how, why, and at what rate new ideas and technology are adopted by populations. Rogers defines diffusion as “the process by which an innovation is communicated through certain channels over time among the members of a social system” (1995, p. 5). According to Rogers, a population is subdivided into five categories depending on their approach and willingness to adopt novelties and Rogers found the distribution across categories to be a normal distribution (figure 1).

![Figure 1: Rogers’ Diffusion of Innovation curve (after Rogers 1995)](image)

According to Rogers, the categories are defined as follows. Innovators (2.5%) always explore something new, no matter if it becomes mainstream or not. Early Adopters (13.5%) wait for the innovators to do the initial work and adopt anything that appears to be useful. Rogers describe Innovators and Early Adopters as individuals who possess an inner motivation to explore and adopt novelties and accordingly they resemble Castells’ self-programmers. Early Majority (34%) adopts something new if it becomes popular among the Early Adopters. Once the Early Majority accepts a novelty, it becomes mainstream. Late Majority Adopters (34%) accepts a novelty because they are expected to follow the general trends of society. They adopt the new because they have to. The last group are the Laggards (16%) who will go to lengths to avoid anything new. Early and Late Majority have to get used to something new and they need to see a purpose. For the Early Majority Adopters the purpose may be a potential they wish to exploit, while Late Majority Adopters accept the new on the basis of a personal cost-benefit-analysis. The fastest Early Majority Adopters may be described as self-programmers while the rest together with Late Majority Adopters and Laggards resembles generic labour as they demand instruction in order to acquire and adopt the new (Levinsen & Sørensen 2008, EVA 2009).

According to Rogers, a person does not belong to the same category in all aspects of life and may even move between categories. The speed of adoption also differs considerably depending on how radical changes the innovation implies in the life of the single adopter or adopting organisation. E.g. WAP-technology did only catch on in Japan while mobile technology spread throughout groups across the world including the laggards in a few years. With this in mind, Rogers’ theory offers a description of aspects of an adopters’ attitude towards innovations that may function as a tool for self-evaluation of the strength of one’s inner motivation for exploring the unknown.

2.1.3 Competence for exploring the unknown

In contrast to Rogers’ theory, Dreyfus and Dreyfus (1988) asked which performative appearances distinguish and define the extremes novice and expert together with the intermediate positions. The Dreyfus brothers studied the practice of chess players and nurses. That is, disciplines that combines formal and informal competencies in terms of rational thinking, profession skills and experiential
knowledge construction. Based on the research and Piaget's constructivist theory they formulated a five stage Model of Skill Acquisition, where each stage describes the phenomenological characteristics of learning strategies for developing expertise.

The novice performs trial-and-error strategies without reflection. The advanced beginner depends on rules but reflects within learning-by-doing strategies. The competent may act independent and deliberately plan and change strategies. However, the competent does not question the basic assumptions. In Piagetian terms, novices, advanced beginners and competent assimilate and accumulate new knowledge. They are - to various degrees - dependent on rules and instruction and they prefer to react rather than to proact. In this sense they resemble generic labour and slow adopters. In contrast, the proficient and the experts act independent, reflective and proactive. The proficients may question basic assumptions and radically change their strategy due to reflection and experience while the expert deals with challenges by intuitively drawing on tacit knowledge and thought experiments in a fluid performance of Reflection in Action (Polanyi 1968, Schön 1983). In Piagetian terms the proficient and experts accommodate new knowledge as they (re)arrange and (re)construct basic assumptions and strategies in radical ways. Therefore they resemble self-programmers or Innovators and Early Adopters. The Model of Skill Acquisition offers a tool to describe how competent one finds ones personal strategies for exploring the unknown to be.

2.2 The content of the cardboard box

![Figure 2: Topobo building elements](image)
Figure 3: An animal creation assembled with Topobo building elements

Topobo is the world's first construction toy with *kinetic memory* = the ability to record and playback physical motion. It is developed by MIT Media Lab's Tangible Media Group (Raffle, Parkes & Ishii 2004).

The prototype consists of passive and active (motorized) components. By snapping together passive and active components, it is possible to assemble dynamic biomorphic forms like animals and skeletons. The system is programmed for movements through direct manipulation of the actives: pulling, pushing and twisting. Apart from individually programmed actives, Topobo also has special actives called "Queens" that control an entire network of individually programmed actives, thus allowing a variety of combinations. After recording movements into the kinetic memory, the user may observe how the assembled construction moves and explore and experiment with complex constructions and movements.

3. Self-programming in practice - the students’ first person perspective

In this section the students’ reflections and learning are presented in a condensed form based on their first person perspective from the written immediate impressions, their e-portfolios and the reflection papers from the course. When working with their e-portfolio and the reflection paper, the students also had access to and actively used the video recordings from the activity.
3.1 Amanda

Amanda was surprised by her reactions. She had expected to act as an Early Adopter according to Rogers. Instead she acted as an Early Majority adopter. In the reflection paper she describes her approach as: *I was open for the new but approached it in a formalized way. I did not expect anything to work.* Amanda had also expected to self-programme on a competent level according to Dreyfus & Dreyfus and found to her surprise that she acted as a novice. She describes how she mimics the other two students or plays around a random in a trial-and-error strategy rather than to reflect on her experience and ask explorative questions. Amanda admits to herself after having struggled, that her dominant barrier is a feeling of impatience that produces a feeling of frustration over the fact, that she is not immediately able to grasp Topobo. Later in workshop I, Amanda decides to exploit her new gained awareness of her strategies to actively improve her self-programming competence. She decides to develop a learning object in Google Maps which is a new application for her. She experiments with ways of posing questions in order to diagnose challenges rather than face obstacles. Amanda concludes that in order to work with her self-programming competence, the task must be relevant and concrete. She also stresses the importance of individual working space along with collaboration and knowledge sharing.

3.2 Brenda

Brenda acts in accordance with her own notion of being Late Majority and lets the others take the lead. She only scratches the surface and plays around with Topobo elements at random. However, looking at the video she realizes that when she suggests something she often drives the collaborative learning process forward. E.g. she is the one to understand the difference between the actives and the "Queen" by just watching the others. Referring to the video Brenda writes: *The teacher (the author) mentioned my suggestions and that my approach seems to be intellectual – that is not at all how I see myself! But the elements sounded like an electromotor and the different colours seemed to behave different. They reminded me about serial- and parallel connections in electric systems in school. I never understood those connections before.* To her surprise she finds herself to be a good observer who reflects and suggests changes in the shared strategy of exploration at Dreyfus & Dreyfus' levels competent and proficient. In the concrete situation the teacher's comment provokes Brenda to change her approach and begins to manipulate the components herself and she invents small experiments that systematically explore gravity and motion in order to construct a forward moving crap together with Amanda. In her professional life Brenda introduces digital support for dyslexic students and helps them to implement the support in their study practice. In the reflection paper she writes that she expects the awareness of her own learning process towards approaching something new, may improve her openness towards dyslexic students' position and reactions regarding the digital support.

3.3 Christian

Before the cardboard box is opened Christian describes himself as an Innovator who likes to fiddle with new gadgets and he expects his approach to self-programming to be proficient according to Dreyfus & Dreyfus. However, he experiences to get stuck with Topobo and ends up in eternal circles of trial-and-error with no progress. In the reflection paper he mentions Brenda's suggestions as a personal turning point that inspired him to change strategy and mimic Brenda: *I left fiddling to the others and began to suggest new procedures and solutions. I found that it is easier to reflect and modify my understanding if I do not always place myself in the first row and fiddle with things.* Later when he looks at the video, he sees how he literally takes things out Amanda's hands. Thus, his urge to fiddle is a barrier not only for himself but also for others. In the next phases of the workshop Christian decides to change his strategy and mix fiddling and reflection in a combination of individual work and collaborative knowledge sharing. He finds this new insight to be important in his professional work as a pedagogical ICT supporter.

4. Discussion

The following images display the change in the participation and the strategies of the students during the day. In the beginning of the day (figure 4) when they open the box, the students' initial roles are observable. In the middle of the day (figure 5) Amanda works together with Brenda while Christian have decided to stop fiddling and observe and reflect instead. In the late afternoon all three have become aware of their barriers and have changed strategies accordingly (figure 6).
Figure 4: This picture from the beginning of the session displays the students' initial roles. Brenda reads the manual; Amanda stays behind with her hands on her back; Christian fiddles.

Figure 5: In the early afternoon both Amanda and Brenda have begun to experiment while Christian have decided to stay behind and observe and reflect on their activities.
In their reflection papers the three students independently conclude, that the workshops design for learning pushes them to reflect on their strategies and patterns when introduced to something new. They express how this experience impact on their awareness of self-programming competence building. Additionally, they find they have improved regarding conscious awareness and the ability to evaluate the appropriateness of a chosen strategy. During the day with Topobo, the students consciously begin to modify their strategies. At the end of the entire module, the impact of change is present and articulated by the students in their final and externally evaluated module assignments. Brenda and Christian claim a direct transfer of learning to their professional jobs while Amanda experience the ability to transfer and exploit the self-programming strategies when she confronts herself with new technological challenges as Google Maps and later other technologies.

From the research point of view, the changes in the students’ behaviour can be interpreted as a movement from a single-strand to a multi-strand strategy. In the beginning the students display the following strategies:

- Amanda is a fiddler but needs a clear purpose or idea before she can manage to reflect on and learn from her hands-on experiences. According to Rogers, Amanda is an Early Majority Adopter and her self-programming approach matches Dreyfus & Dreyfus’ novice level.
- Brenda observes and avoids fiddling. She pushes the responsibility of action away from herself and acts as Rogers’ Late Majority as she rarely explores anything entirely new. However, her intellectual self-programming approach matches Dreyfus & Dreyfus’ competent to proficient levels.
- Christian is a fiddler and attacks any gadget that comes his way. He is obviously an Innovator when it comes to technology. However he is at Dreyfus’ and Dreyfus’ novice level when it comes to self-programming.

The strategies are all single-stranded; they either fiddle (Amanda, Christian) or reflect (Brenda). When the students reflect on their actions and attitudes, they find that although their preferred strategy seems to work, they all encounter impenetrable barriers. In their efforts to pass the barriers, the three
students independently realize that fiddling has to be combined with reflection and they gradually move towards multi-strand approaches. In acquiring a multi-strand approach the students gradually become able to consciously change the initial basic assumptions about Topobo and accordingly their strategies towards Topobo. That is, they move their competence level of self-programming towards the proficient level.

In the process of becoming aware of self-programming competence building as a personal learning process, Topobo seems to play two parts. In the individual space the student gradually figures out about Topobo and becomes aware of self-programming – that is learning as a process becomes observable and thereby also an object of change for the student. Topobo can be said to support a cycle of internalization of what is outside, and externalization of what is inside the mind of the single student. This process corresponds with Seymour Papert’s ideas of constructionism (Papert 1990, p. 3). In the collaborative space the externalized ideas that materialize in specific Topobo constructions function as a boundary object (Wenger 2000). The materialized construction represents an inexpressible idea or tacit knowledge (Polanyi 1968) as Brenda’s sudden flash back to serial- and parallel connections in electric systems. In this sense, the construction as a boundary object becomes a non-verbal language that supports a shared negotiation of meaning and a shared re-arrangement of materiality. During this process the students becomes aware of how their strategies either promote or hamper the shared exploration of Topobo.

Gradually, the multi-strand strategies come closer to a bricoleur-strategy. According to Turkle and Papert a bricoleurs approach to challenges is to connect practice and concrete thinking in an intertwined process of arrangement and re-arrangement of materiality while constantly negotiating and re-negotiating meaning (Turkle & Papert 1990, pp. 129). This is in accordance with Amanda and Christian who both stresses the need for individual space together with a collaborative and social space for learning. In contrast, Brenda primarily stresses the collaborative space as she prefers to be instructed either in a formal context or by fellow students. This finding is important, as a dominant trend in current and future oriented designs for teaching and learning emphasizes collaboration and social learning at the expense of the individual learning space.

In the reflection papers, the students individually conclude that self-programming competence building must evolve round a task with a meaningful outcome. However, meaningful means something different depending on the students’ willingness to explore and adopt new technology. This is an important finding regarding how to design for self-programming competence building, as “on-size-fitt’s-all”-design may prove to be insufficient. For Christian as an Innovator it is meaningful to explore out of pure curiosity, while Amanda as an Early Majority Adopter explores technology when she has seen examples of use that inspires her to imagine learning potentials in her own field. Brenda as a Late Majority Adopter only explores new technology when she is forced to do so. However, during the workshop she gradually becomes an Early Majority Adopter.

The findings from the case study are both related to the individual and collaborative barriers and emerging proactive strategies that come into play. Drawing on these findings, it is argued that the presented Topobo-session design contributes to the networked society’s design for teaching and learning, as the design – at least for this small group of students - supports their development of self-programming as a sustainable competence. In the autumn 2010 the study will be expanded to a larger group of students.

References
Discovering Student web Usage Profiles Using Markov Chains

Alice Marques and Orlando Belo
University of Minho, Portugal
alicemarques5@gmail.com
obelo@di.uminho.pt

Abstract: Nowadays, Web based platforms are quite common in any university, supporting a very diversified set of applications and services. Ranging from personal management to student evaluation processes, Web based platforms are doing a great job providing a very flexible way of working, promote student enrolment, and making access to academic information simple and in an universal way. Students can do their regular tasks anywhere, anytime. Sooner or latter, it was expected that organizations, and universities in particular, begin to think and act towards better educational platforms, more user-friendly and effective, where students find easily what they search about a specific topic or subject. Profiling is one of the several techniques that we can use to discover what students use to do, by establishing their user navigation patterns on Web based platforms, and knowing better how they explore and search the sites’ pages that they visit. With these profiles Web based platforms administrators can personalize sites according with the preferences and behaviour of the students, promoting easy navigation functionalities and better abilities to response to their needs. In this article we will present the application of Markov chains in the establishment of such profiles for a target eLearning oriented Web site, presenting the system we implemented and its functionalities to do that, as well describing the entire process of discovering student profiles on an eLearning Web based platform.

Keywords: web based elearning platforms, web usage profiling, Clickstream analysis, Markov chains, Navigation paths analysis

1. Introduction

The development of a web site by an organization may not be enough to obtain the success it expect on the Internet. Being the web a market with a large facility to publish contents, certainly will exists other sites, with the same or very similar contents, that users can use as an alternative. Thus, it’s necessary to ensure that the site is in line having what its potential users want, in order to avoid they abandon it after some time of using, going to another site with better contents proposals, a more adjusted organization to their needs, or incorporating strong user friendly navigation facilities.

During the last few years, the number of students using web based platforms for learning activities grew a lot, reaching a level quite interesting to analyse the way how and when they use those platforms and their resources. Goals are very clear: to optimize quality of service, to increase resources availability, to attenuate the effect of services (and resources) not used, and to facilitate access to didactic material 24 hours a day. Today, we see universities promoting the development and maintenance of very effective sites, being their administrators quite concerned about their use and effectiveness. Some of them use to monitoring the activities performed on their sites, especially the ones that support eLearning platforms, in order to provide better quality of service, to supply better information (on time) and to avoid eventual system’s downtimes. Additionally, this way of looking for web usage has been improved in the sense to discover critical periods of site navigation and exploitation, trying to prevent service bottlenecks and lower rates of service quality – these are two of the most critical aspects that lead students to avoid or abandon eLearning oriented Web sites.

For a university site be successful, we believe we must know, as in traditional business areas, the type of users that navigate in the site, once they are the primordial elements that will influence site advancement, as well as the evolution of the organization itself. That is why more and more organizations try analyzing several questions, related with the form how users interact with the site - which pages are more and less visualized, what are the main pages of entry and exit the site, which are the hours with more traffic, the frequency with which a user visits the site, and so on. The analysis of this type of questions will identify student usage profiles that use a particular eLearning oriented site. The work of identifying usage profiles is a systematic activity, involving massive exploitation of all the available clickstreams, storing, transforming and analysing them with the most recent techniques to do that. With these profiles established we could personalize a site according with the needs of students, promoting easy navigation and better ability to response to customers needs.

To establish student profiles, identifying their behaviour when using a eLearning site, we need to monitoring their daily usage and collect all information we can about their activities inside the site,
knowing what kind of links they use to follow, services or documents they use to access, their frequency of usage, or simply what are their usual entry points. All of this can be done observing the access logs of the site, and analysing things such as the IP number of the machine used during the navigation process, the pages visited (their links, and data and time of access), time of permanence in a particular page, and the resources used, just to name a few. Taking as the IP address as the basis to characterize student navigation, we can ordered the pages by date and time of access and group them into sessions, considering a predefined navigation period, which will define the period to group the visited pages, and a period of inactivity that will define the end of a particular session. Then, analysing all the student sessions we can establish navigational patterns, representing the most regular paths that students use to follow during a session, and consequently generate a profile to personalize (if necessary) the contents of a site.

In this paper we will present the application of Markov chains in the establishment of usage profiles for a target eLearning oriented Web site. With the profiles established we can identify the type of user that is visited and with this knowledge we can provide to students in a better way the services that they probably will use in a near future. Additionally, we will present the way we collected the data, designed and explored the mining models, and the way that Markov chains conducd us to the establishment of the student web usage profiles. We finalize the paper with some conclusions and presenting some paths for future work.

2. Using web based eLearning platforms

2.1 Overview and general features

Web based platforms are expanding their influence in many sectors. From services to industrial plants we can find several applications in the area. ELeaning (Carliner & Shank 2008), as we know, it is not an exception. For many years educational institutions began to install and explore Web based platforms, frequently only as a simple way to promote themselves or to receive documents repositories. Quickly they discover the huge attractively and flexibility of Web based platforms, a great freedom in courses’ contents management and maintenance, and as a practical way to improve learning processes, share knowledge and expertise, augment student enrolment, and of course reduce operational and maintenance costs.

With the emmergence of Web based eLearning platforms new horizons has been open to the implementation of new studying environments. These platforms help educational institutions training diverse student populations and provide a very attractive way to enlarge classrooms, taking them where students are. Today this is not a novelty it is a clear reality. The improvement of the Internet and its services helped a lot on this process. Its universal access and its easy usage facilitate a lot the adoption of these platforms. It is a market in a continuous expansion, and it is easy to see that for the number of educational tools that appear every single year. Today, the difficulty is not their use and exploitation but its selection and adoption. Universities have now a lot of options concerning web based educational platforms. Model1, Blackboard2, Claroline3 or Atutor4 are only a few references that we can find in the market. The services they provide are quite diversified. All of them work towards helping educators to create, manage and maintain online courses, and all their related services, for large communities of students, having abilities to cover educational topics from primary schools to universities. To do their jobs well, Web-based eLearning platforms provide a set of features very powerful, which includes today things such as integration of multimedia objects, multilingual support, project management tools, data import and export services, personalised access based on role definitions, activities reports, evaluation tools, or heterogeneous document types hosting, just to name a few (Tucker et al. 2002). There is an evidence that Web-based ELearning systems (Schewe et al. 2005) are a clear reality with a significant impact in our lives. We face them practically in any educational institution supporting current activities and adding new value to personal education and student enrolment (Hosan et al. 2006). Day after day, educational managers give more attention to these platforms stimulating and supporting their design based on student profiles, and creating flexible and friendly navigation structures, as well are creating methods and processes to model new Web based educational systems (Rokou et al. 2004). Profiling is one of the best ways we have to go towards an adaptive Web-based eLearning system.

1 http://moodle.org/  
2 http://www.blackboard.com/  
3 http://www.claroline.net/  
4 http://www.atutor.ca/
2.2 Improving eLearning web platforms through profiling

As their experience grows on using Web based eLearning platforms, educational institutions feel the need to improve their sites and related services. Reasons could be quite diverse. However, identifying, understanding and characterizing trends in what users do when they are navigating on a specific Web based eLearning site is with no doubt one of the most important reasons why institutions (and their webmasters and educational managers) are doing Web profiling nowadays (Lourenço & Belo 2009). Students using such platforms are a little bit different from the ones that access to non-educational Web sites. But, the way of acting is the same, as well are the tools and the services available. Thus, we can apply over Web based eLearning sites the same techniques that are currently used over any other site, independently from its business area, to catch the way of access, being, and explore a web site.

Web profiling (Spiliopoulou et al. 2000) is today one of the leading research and technical area of Web usage analysis (or clickstream analysis). The establishment of Web usage profiles allow us to evaluate the effectiveness of a site’s contents, to improve their services performance and, consequently, to assess its popularity and effective use, which could help restructuring Web sites towards a more suitable platform for students and, of course, for their promoters.

There are several methods and techniques to improve Web sites organization and contents that we can use. Web usage analysis (Borges & Levene 1999) provides us the means to cover all the essential phases to discover a Web profile based on the page views request sequences records (clickstream data) we have available, namely (Ramadh an et al. 2005): pre-processing, where we read and interpret clickstream data in order to identify (if possible) an user (or a group of users) and its sessions, giving particular attention to information about applications, content server, and visited links and associated duration times; pattern discovery, in which we try to evaluate several functional and operational indicators such as sites more visited, sessions average lengths, IP addresses resolution, users’ countries; most used agents, page views more frequent and respective access times, or navigations paths; and, finally, pattern analysis, where we deal with very specialized tasks that distinguish relevant information in clickstream data for user characterization. Classification and clustering, conventional statistics analysis and sophisticated visualization techniques tools are frequently used in the most Web usage analysis processes, helping a lot in the establishment of Web usage profiles. However, other techniques and models to identify profiles could be used.

3. Profiling with Markov chains

In Web Usage Mining (Srivastava et al. 2000), sequential patterns techniques (Esmaeili & Gabor 2010) try to find regular occurrences of a same set of elements (patterns) for each Web session under supervision, such as the presence of an item set followed by other item in a target group of page views. Using these techniques it is possible to predict future visit patterns that could be used to launch oriented advertisements, functional warnings, or simply to inform Web users about new things that they could be interested. Basically, we intend to find some relevant information about Web users behaviour, in order to provide more attractive site. A usual navigation process over an ordinary Web site begins by selecting an initial hyperlink and receiving the page view within a conventional Web browser. This first page view defines the initial state of a new Web session. Next, the navigation process continues selecting a new hyperlink that will define a new stage for the current Web session. This successive selection of hyperlinks is continuously recorded, click after click, in specific clickstream files. Latter, and with appropriated techniques, the analyses of these log files will reveal all the navigation paths (sequences of pages views) for a Web site that its users have done their sessions (Jansen et al. 2007).

As referred before, one of the most relevant approaches to map Web navigation paths is to use Markov chains (Sarukkai 2000). They were introduced as a mechanism to predict potential Web navigation links in a Web site (Zhu et al 2002) (Deshpande & Karypis 2004). With a Markov chain it is possible to indicate which is the next page will be requested by an user based on its current location and on its previous navigation sessions. Also, when generated based on Web transactions, the chains provide us the navigation paths followed by a user during a specific period of time, giving us the possibility to identify the most frequent page sequence that the user will probably follow in a future Web session (Borges & Levene 1999). Representing Web navigation paths with Markov chains we can get valuable information about the site in analysis and its future users navigation tendencies (Mobasher 2006) – for instance, through the analysis of a Markov chain, it is possible to predict what
pages view sequence a special kind of user will do in a newspaper, and prepare accordingly some advertisement spots that go towards the characterization of the navigation profile.

Markov chains are especially useful to built prediction models (Deshpande & Karypis 2002) (Surukkai 2000), allowing for the establishment of future user behaviour while users are interacting with the sites. This is done with the analysis of previously users behaviour with similar interests. We can use Markov models to find the more frequent trails (navigation paths) followed by users in their navigation processes, which means to find the most frequent sequences of pages that the users visit during their navigation sessions. Markov chains could be seen as a conventional graph in which nodes (stages) represent visited pages and edges (transitions) the probability related to some Web user passing from a page to another. Usually, edges' probabilities are calculated counting de number of users that pass from one page to another, taking into consideration the number of visits for each origin page. More formally, a Markov chain is characterized by a set of states \( \{s_1, s_2, \ldots, s_n\} \) and a matrix of probabilities \( [P_{ij}]_{nxm} \), where \( P_{ij} \) represents the probability to move from state \( i \) to state \( j \).

The Markov chains are especially useful for predicting models based in continuous sequences of events – their application is quite direct to Web sessions. The order of a Markov model represents the number of prior states used to prevent the next state. So, a Markov model of order \( k \) predicts the probability of the next page based on the last \( k \) pages visited. Given a set of all trails \( R \), the probability of reaching a state \( s_i \) from a state \( s_i \), via a trail \( r \in R \), is given by \( \Pr(r) = \prod_{j=k+1} r_{ij} \), where \( k \) ranges between \( i \) and \( j-1 \), in other words the probability is given by the multiplication of all intermediate states (Mobasher 2005). As an example about how a markov chain can model a set of web transactions, consider the set of transactions presented in Table 1.

### Table 1: An example of a transaction set

<table>
<thead>
<tr>
<th>Id</th>
<th>Transactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A → B → C → D → E</td>
</tr>
<tr>
<td>2</td>
<td>A → B → C</td>
</tr>
<tr>
<td>3</td>
<td>A → B → C → E</td>
</tr>
<tr>
<td>4</td>
<td>C → D → E</td>
</tr>
<tr>
<td>5</td>
<td>C → D → E → B</td>
</tr>
<tr>
<td>6</td>
<td>C → D → A → E</td>
</tr>
<tr>
<td>7</td>
<td>D → A → B → E</td>
</tr>
</tbody>
</table>

To build a Markov chain we start by adding an initial state (S) into the chain. This state will have a transaction for all the web pages visited for the target web site. Then we define a final state (F) that will be the end for every last page of either session. This means that every last page visited will have a transaction to this final state. The probabilities associated with the edges are obtained by counting the number of times that the transaction occurs in the trails. Thus, the probability to move from the initial state S to a state A that represents the page A is about 5/27 (0.18), where 5 is the number of times that the page A occurs, and 27 is the total number of requests. Using the same process, the probability to move from page A to page B is 4/5 (0.8), where 4 is the number of times that B occurs after A, and 5 is the number of times that A occurs. Finally, the probability to move from page E to the final state F is 5/6 (0.83), where 5 is the number of trails where E is the final state, and 6 is the number of times that E occurs. The Markov model generated from such transactions is depicted in figure 1.

However, a Markov chain it is not enough for a correct identification of a Web user profile. It is also necessary to calculate which are the most frequent Web paths (a sequence of Web links) followed on a specific site by its users. This is usually done after the Markov chain was built. To do that, we can use a common search algorithm for graphs, such as the Breath-First Search (Weiss 1993) or the Depth-First Search (Tarjan 1962) algorithms.

In order to calculate the frequent trails (table 2) we have to take into consideration two very important concepts: support and confidence. The support is given by the initial probability of the chain, which is frequently defined as the average of all the initial percentages of the Web pages, and the confidence by multiplying the probabilities of all links into the chain. The frequent trails will be calculated as follow: first we get all pages that have initial probability higher then the support value; then we use the
depth-first (Tarjan 1972) algorithm to search in graphs, trying to find the tails that have a probability bigger than the confidence.

![Diagram](image)

**Figure 1:** An example of a Markov Chain

**Table 2:** Frequent trails with support = 0.1 and confidence = 0.4

<table>
<thead>
<tr>
<th>Trail</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>B -&gt; C</td>
<td>0.6</td>
</tr>
<tr>
<td>B -&gt; C -&gt; D</td>
<td>0.4</td>
</tr>
<tr>
<td>D -&gt; A</td>
<td>0.4</td>
</tr>
<tr>
<td>D -&gt; E</td>
<td>0.6</td>
</tr>
<tr>
<td>A -&gt; B</td>
<td>0.8</td>
</tr>
<tr>
<td>A -&gt; B -&gt; C</td>
<td>0.48</td>
</tr>
<tr>
<td>C -&gt; D</td>
<td>0.67</td>
</tr>
<tr>
<td>C -&gt; D -&gt; E</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Finally, analysing each identified path and using their own knowledge and expertise about the site itself, we can easily to establish the site’s Web usage profiles, simply because paths characterize quite well the type of users that visit the site. These are only some general ideas about Markov chains and their utility in the establishment of Web usage profile - to have a more detailed idea about this we suggest the reading of (Sarukkai 2000), (Zhu et al 2002) or (Deshpande & Karypis 2004).

4. Establishing student web usage profiles

4.1 The eLearning site target

In order to demonstrate the utility of Markov chains in the establishment of user usage profiles in eLearning platforms, we selected one of our educational sites to be our target of study. This site (Figure 2) is supported by Moodle (Cole & Foster 2008) and maintains all the information about a MSc curricular unit that students need to be update about curricular topics, analyse and discuss practical issues, and receive or post their practical works. The eLearning site was structured to receive several courses. From its initial page, valid credential users can select and access to the material posted for each course on-line. The platform is able to support the main operations that ones need to use a site like this, namely: consult, insert, update or delete information items for all the resources available, communicate and exchange information with other students, follow the execution of practical works,
see evaluation processes, access and read technical reports and scientific papers, or participate on discussion areas. Additionally, we can also use some other management services that provide us generic information about current users and their activities.

The target public of our site are teachers and students of the courses the site maintain. As expected, the student community is the largest one. It is quite diverse and heterogeneous, and can be divided in different groups accordingly their own navigation credentials. Which means that it is a good target for our first studying process for discovering and establishing eLearning usage profiles.

![Figure 2: The target eLearning site](image)

The site doesn’t have a large volume of accesses, once it was designed and implemented for courses with small number of students. A preliminary site usage analysis gave us some curious information, but nothing very surprisingly to people that know a little bit about educational institutions and their working periods. Every time a course begins the site receive a lot of access requests and page viewing increases quite much when compared to other periods - students are looking for information about the functioning of the course, and preliminaries documents with course’s program contents and evaluation criteria. A same high usage level happens again, every time one of the teachers post the results of a evaluation process or launch new tasks to do in the context of a specific working plan. These mean a very simple thing: every time students detect a new piece of information on the eLearning site they access it. Finally, and as expected, the majority of the accesses to our site were done from Portugal, once the students are Portuguese and every time they decide to visit the site use Portuguese access points.

### Table 3: A simplified view of an event log file structure

<table>
<thead>
<tr>
<th>Field</th>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course</td>
<td>Mi_SSD_1, 2009 Dezembro 10 19:19, 188.81.39.207, userName, course view, Mi-SSD 2009/2010</td>
<td>Course identification.</td>
</tr>
<tr>
<td>Date</td>
<td>2009 Dezembro 10 19:19</td>
<td>Date and time of the Web server when responding to a service request.</td>
</tr>
<tr>
<td>Host</td>
<td>188.81.39.207</td>
<td>User’s IP address.</td>
</tr>
<tr>
<td>User</td>
<td>UserName</td>
<td>User name.</td>
</tr>
<tr>
<td>Page</td>
<td>Course view</td>
<td>Service requested.</td>
</tr>
<tr>
<td>Information</td>
<td>Mi-SSD 2009/2010</td>
<td>Information about the service requested.</td>
</tr>
</tbody>
</table>
4.2 Sources and data preparation

All target data was collected in the information sources used by Moodle to identify and characterize students, and record their activities on site in a specific event log file. These log files are quite important to understand user behaviour. Their structure (Table 3) allows keeping detailed records about tasks users performed, including also date and time stamps and resources accessed.

Data preparation is one of the most important steps in profile identification and description. It requires that data will be transformed and clean according with the requirements of the data mining techniques to be used. As usual in these cases, to prepare data for profiling we need to perform several tasks that goes from users identification to data conciliation, passing for other tasks such as session delimitation, data cleaning and enrichment or IP addresses resolution, just to name a few.

We started the data preparation process by removing from the log file every line (record) that contained one or more null values in relevant attributes used on the users sessions reconstruction, because nulls do not let identify any kind of profile – as we know, a null value represents an unknown value and acts in any process as an absorbent element. Next we passed to the identification of the users that visit our site. This is quite important if we want to characterize their behaviour and consequently their usage profile. There are several methods we can use to make this identification (Cooley et al. 1999). However, in our case, this process was very simplified, once our target site demands that users identify themselves before access to the eLearning platform. So, when they do that, the system records their identification for any task they perform inside it. The next step was user session identification.

![Figure 3: The user session star-schema](image)

Basically, a session is a set of actions that a user performs in a site from the moment he enters it to the one he leaves it (W3C 1999). User session reconstruction intends to group under a unique identifier all the page views requests that a user did for a site. Keeping this kind of information (clickstream data) it is possible to study users’ behaviour as well to know their navigation preferences and tendencies. We can divide session reconstruction techniques in two categories: proactive and reactive (Spiliopoulou et al. 2003). The first one considers that the session unique identifier is assigned during the navigation process, whilst the second defines it only when the session ends, appealing to the data recorded in the navigation logs. We used a reactive technique in our case study.
So, we began to define a maximum period of inactivity: 10 minutes. Next, we organized all page views requests by date and hour for each user registered in the site, in order to calculate user sessions. If two successive requests were posted in a time interval less than the maximum period of inactivity they belong to the same session. In the contrary case, we created a new session.

The last step in the data preparation process was loading all the prepared data in a data warehouse, which was especially designed and implemented according the requests for usage profile identification and characterization. The star-schemas organization followed the one presented in (Kimball & Merz 2000) for similar cases. We stored the data in two different star-schemas: one for receiving page views requests, and another to keep information about user sessions (figure 3). In this last star-schema we easily identify five dimension tables: 1) calendar (time_dim), which defines the different time periods (day_of_week, month, quarter, year, etc.) for user session analysis; 2) hour (hour_dim), corresponding to the different hours of the day and correspondent periods; 3) web pages (pages_dim) that contains the hyperlinks of every single page that was visited by users; 4) users (user_dim), which has a short identification of the user (only possible because all eLearning platform’s services require access credentials); and curricular units (curricularUnit_dim), the dimension that has all the names of the curricular units supported by the platform. In order to define the N:N (many-to-many) relationship between the web pages dimension and the user session fact table (TF_Sessions) – one single Web session usually include several Web pages, and a Web page could be found in several distinct Web sessions - we created a bridge table (TP_Session). A quick look to this multidimensional schema allows us to see the enormous potential that a data structure likes this one has, and the enormous amount of queries that it is possible to satisfy about user sessions. For instance, if we want to know the average session time of all users that accessed the ‘Computer Science’ curricular unit site every ‘Friday’ during ‘2010’, we can launch, in a simple manner, the following SQL query:

```
SELECT US.Name, AVG(TF.Session_Duration)
FROM TF_Sessions AS TF INNER JOIN Time_Dim AS CA
    ON TF.StartTime = CA.Time_Id
    INNER JOIN User_Dim AS US
        ON TF.[User] = US.Id
    INNER JOIN CurricularUnit_Dim AS CD
        ON TF.CurricularUnit = CD.Id
WHERE CA.Day_of_Week = 'Friday' AND CA.Year = '2010'
    AND CD.Name = 'Computer Science'
GROUP BY US.Name;
```

or to know the 10 most used curricular unit sites:

```
SELECT TOP 10 CD.Name, SUM(TF.Session_Duration)
FROM TF_Sessions AS TF INNER JOIN CurricularUnit_Dim AS CD
    ON TF.CurricularUnit = CD.Id
GROUP BY CD.Name
ORDER BY SUM(TF.Session_Duration) DESC
```

or, finally, to get the most accessed Web page in all curricular unit sites:

```
SELECT TOP 1 PD.Page, COUNT(TF.Session_Id)
FROM TF_Sessions AS TF INNER JOIN TP_Session AS BT
    ON TF.SessionId = BT.Session_Id
    INNER Pages_Dim AS PD
        ON BT.Page_Id = BT.Id
GROUP BY PD.Page
ORDER BY COUNT(TF.Session_Id) DESC;
```

In this work we only used the data loaded in the user sessions star-schema, corresponding to 501 sessions performed by users between ‘2009/12/10’ and ‘2010/05/26’ and generated from a data set of 3585 requests. These requests involved 25 distinct users that accessed to 41 different pages. All this corresponds to the activity developed in an eLearning system for a single course integrated in a Decision Support Systems curricular unit. In spite of being a relative small data set, it was enough to
demonstrate the utility of Markov chains in the establishment of usage profiles – an excerpt of the contents of the user session fact table can be seen in figure 4.

4.3 The profiling process

After data preparation and integration, the next step on the profiling process is the generation of the Markov chains. This task begins with the reconstruction of the site’s users sessions, which are necessary to support and represent the chain. To do that, we collected all the pages that were visited, and retrieved all the requests done over them, ordered by session and sequence number. With all the sessions reconstructed and all the visited pages stored, we calculated the initial probability for each visited page, dividing the number of requests for the page by the total number of requests. Having all the initial probabilities set, we must to calculate the probability that a page have to appear (be accessed) next to other.

Figure 4: A record set fragment of the user session fact table

To do this, we need to get again all the pages visited, and page-by-page get all the sessions where each of them appears. ENDING that, it was necessary to find what page appears next to the previous referred page in all sessions. If that happens for a first time we initialize a counter with 1; if not, we increment the counter by 1. At the end we have an array with all the pages next to the reference page and the number of times they appeared in that position. The probability of the next page is calculated taking the number of times that a page appeared next to the reference page and dividing it by the number of times that the reference page appears. In Figure 5 it is presented an excerpt of the Markov chain generated for the site, using the tool GraphViz due to the large number of nodes and edges of the Markov chain generated, we only present part of it in this paper.

Figure 5: An excerpt of the Markov chain generated for the site

---

http://www.graphviz.org/
After the generation of the Markov chain and getting the user accesses data, we calculated the most frequent paths followed by users. Then, we defined the minimum support, using the average value of all the initial probabilities defined previously, and the confidence, selecting a value of 15% due to the fact that users had a large number of pages for selection. If we had chosen a greater value for the confidence, an important number of chains simply was despised and not included in the final results. To get the most frequent paths we need to: retrieve all the pages views that have a probability greater than the support that was defined; calculate all the paths whose confidence is greater that the defined value using the depth-first algorithm for graphs (Tarjan 1962); and, finally, show a list (Table 3) of the most frequent paths with a support of 1% and a confidence of 15%. At this time, these are two important concepts to consider - support, represents the initial provability of a Web path, many times defined as the average of the initial percentages of all the Web pages; and the confidence, corresponds to the probability of an user going on a specific path, being calculated by multiplying the probabilities of each chain connections.

Through the analysis of the most frequents paths that were detected it was possible to identify some relevant site usage profiles. At the beginning, we had made the identification of two distinct groups of users that access regularly to the Web based eLearning site: teachers and students. It was a simple observation of a fact. However, if we look now to the rules that were generated (Table 3), we can see that paths 1, 7, 12 and 19 are associated with the teacher profile, since they include actions related to the creation of new tasks on the site, which are obviously from its competence. Looking for the other rules, we can say that they are typical from a student usage profile. Rules 3, 4, 5, 7, 10, 14, 15, 16, 20 and 22 reveal a student profile a little bit more specific. This first student profile corresponds to a group of users that only visit the site to see information about their curricular units and the tasks they have to do. There is a second student profile that was identified and rules number 6, 11,13 and 18 support it corresponding to users that do the same as the previous profile but also visit the pages from other users defined in the site. Finally, rules 8, 9, 18, 20 and 21 reveal a third students group that are more active on the site, participating in the discussion lists besides other regular activities of studying and working tasks.

Table 3: A list of the most frequent paths

<table>
<thead>
<tr>
<th>Id</th>
<th>Path</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>assignmentupload -&gt; assignmentview</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>assignmentviewsubmission -&gt; assignmentupdategrades</td>
<td>63</td>
</tr>
<tr>
<td>3</td>
<td>assignmentviewwall -&gt; assignmentview</td>
<td>56</td>
</tr>
<tr>
<td>4</td>
<td>resourceviewall -&gt; courseview</td>
<td>44</td>
</tr>
<tr>
<td>5</td>
<td>Courseview -&gt; assignmentview</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>userviewall -&gt; courseview</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>resourceviewwall -&gt; resourceview</td>
<td>34</td>
</tr>
<tr>
<td>8</td>
<td>forumviewdiscussion -&gt; courseview</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>forumviewforum -&gt; courseview</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>resourceview -&gt; assignmentview</td>
<td>30</td>
</tr>
<tr>
<td>11</td>
<td>coursercent -&gt; userview</td>
<td>29</td>
</tr>
<tr>
<td>12</td>
<td>assignmentview -&gt; assignmentupload</td>
<td>25</td>
</tr>
<tr>
<td>13</td>
<td>userviewall -&gt; userview</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td>Courseview -&gt; resourceview</td>
<td>24</td>
</tr>
<tr>
<td>15</td>
<td>resourceview -&gt; courseview</td>
<td>23</td>
</tr>
<tr>
<td>16</td>
<td>resourceview -&gt; courseview</td>
<td>23</td>
</tr>
<tr>
<td>17</td>
<td>coursercent -&gt; assignmentview</td>
<td>21</td>
</tr>
<tr>
<td>18</td>
<td>userview -&gt; courseview</td>
<td>21</td>
</tr>
<tr>
<td>19</td>
<td>forumviewdiscussion -&gt; userview</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>assignmentupload -&gt; assignmentview -&gt; courseview</td>
<td>16</td>
</tr>
<tr>
<td>21</td>
<td>resourceviewwall -&gt; courseview -&gt; assignmentview</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>forumviewforum -&gt; forumviewdiscussion</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>assignmentview -&gt; courseview</td>
<td>16</td>
</tr>
</tbody>
</table>

(…)

©Academic Publishing International Ltd
5. Conclusions and future work

In this paper it was presented a first attempt to establish valid and useful students usage profiles for a Web-based eLearning system of an educational institution. As referred previously, the main goal was to understand how students used the resources that teachers made available to them in an eLearning site. Nowadays, the success of any eLearning platform is directly dependent from the knowledge that its administrator has about their users. Usage profiles can be use by administrators to personalize contents and services towards their users’ needs and expectations. We can use several data mining techniques to identify usage profiles like clusters, association rules, sequential patterns, classification, and so on. The goal of this paper was to study the application of Markov chain to identify student usage profiles of a specific Web based eLearning platform. In general terms, we started by a brief explanation about how Markov chains can be used to discovery web usage profiles, and then we present a case of study where we used Markov chains as a way to identify and establish usage profiles. The intention was to get valid profile information in order to optimize the site's structure and reduce operational costs involved with the maintenance of resources not used. The results obtained were quite typical for a system like the selected one. We identified two large user groups (teachers and students) whose behaviour was divided, respectively, doing creation and maintenance tasks over the site’s contents, and consulting working tasks or participating in discussion lists. Doing this, it was possible to demonstrate the establishment of real usage profiles of the target site through the use of Markov chains. If we had a more significant amount of log records, results will be more remarkable revealing other potential profiles beyond the ones that we obviously expect. This will be explored soon in one of our research lines, especially oriented to the refinement of the generate Markov chains, filtering irrelevant access to some pages in order to discover the most accessed sites and correspondent visit times. This will improve our perception about eLearning sites usage, contributing to improve sites’ organization, contents, interface and to reduce computational costs related to pages that never will be consulted.

References

Assessing Student Transitions in an Online Learning Environment

Minoru Nakayama and Hiroh Yamamoto
CRADLE, Tokyo Institute of Technology, Tokyo, Japan
nakayama@cradle.titech.ac.jp

Abstract: Assessment surveys of students are often conducted in order to evaluate online learning activities. Most surveys measure responses to questions which are based on students’ subjective impressions. The purpose of this study is to examine participants’ assessments made during the transitional phase in an online learning environment which includes blended and fully online courses at a Japanese national university. Students were enrolled in two-unit Master’s or Bachelor’s degree courses which were taught by the same professor. The total number of students with valid survey data was 184 (92 Masters, 67 Bachelors for the blended learning course and 25 Bachelors for the fully online course). A survey questionnaire consisting of 10 questions measured the self-assessments of students’ online learning experiences. Three factors were extracted. There are no significant differences in all factor scores between the beginnings and the ends of the courses. These results show the coherence of students’ assessments during the course. The correlation coefficients of the first factor scores (e-learning evaluation) between the beginnings and the ends of the courses are not high, however (Masters: r=0.35, Bachelors for blended learning: r=0.46, and Bachelors for fully online: r=0.33). Therefore, some participants have changed their evaluations between the two surveys. When the differences in factor scores from the initial and final surveys are compared between students who rated the course highly at the beginning (high raters) and students who did not (low raters), the scores for the high raters decrease and the scores for the low raters increase. Also, the relationships between students’ transitions and the metrics of their behaviour were investigated.

Keywords: online learning, student assessment, assessment in transition, blended learning, fully online learning

1. Introduction

The use of online learning environment is spreading widely in the curricula of modern university courses, in accordance with the development of information technology and online communications (Twigg 2001). When these technologies are introduced into the learning environment, learning performance and effectiveness can be discussed as a cost benefit (Bates 1999). Since it is well known that it is not easy to emphasize the cost benefit of a new technology or a new educational system, determining the impressions and level of satisfaction of stake-holders is often preferred to surveying participants’ opinions, according to the first stage of the Kirkpatrick model which suggests that they be observed and measured (Watkins et al. 1998).

Therefore, assessment surveys of students are usually conducted in order to evaluate progress in the learning environment. Most surveys measure responses which are based on a students’ subjective impressions. In particular, as the online learning environment is very different from the conventional learning environment, both students and teachers have to adapt their abilities. These assessments help to improve the teaching material, course content and supporting methodologies (Harrington & Reasons 2005; Anderson et al. 2006). Additionally, learning skills and student's attitudes should be observed throughout the learning process, and appropriate survey items should be developed to help design support programs for both students and teachers (Craig et al. 2008). Recently, these evaluation results have been used formally, for the purpose of assuring the quality of instruction. The role of assessment has been well recognized as an important management activity (Deepwell 2007). In these scenarios, both students’ satisfaction surveys and assessing the level of student's academic performance are major indices, and can be used as measures of teaching quality at universities (Elton 1993).

Also, the issue of quality assurance is becoming more important in Japanese higher education, as teaching quality and improvements in methodology are major topics of concern for global universities (Marginson 2006). To provide learning opportunities for university students, the online learning environment is an appropriate tool (Kaneko 2009). The differences between students’ and teachers' recognition of their roles in online education were measured over two years using an annual survey (Palmer and Holt 2009). According to the experiences of students and teachers, the results of students’ evaluations have changed across three years (Nakayama et al. 2009).

As academic terms are not short, student's perceptions and skills of adaptation may change during the course. Their attitudes toward the learning environment may develop week by week, as conventional psychological phenomena such as the primary effect and the Hawthorne effect are often observed.
(Haebara et al. 2001). The subjective evaluations of the learning environment by those in a distance education system were significantly different between the initial and final class sessions (Shimizu and Maesako 1988). The online learning environment is evaluated using factors which include material evaluations and self assessment (Nakayama et al. 2006); these evaluations may change along with the progress of learning, and also may be affected by various metrics, such as the number of class sessions attended. These transitions in subjective evaluations may be related to learning performance and other metrics of students’ characteristics, which were extracted during previously conducted studies (Nakayama et al. 2008, 2009).

The purpose of this study is to examine participants’ transitional phase assessments in a blended learning environment and in a fully online learning environment at a Japanese national university. Also, the effectiveness of these transitions on learning performance is investigated. This is a small scale case study, however, because courses differ from university to university. In this study, the determination of the student’s change in evaluation during a course is focused on initially, using conventional statistical analysis of the survey responses.

2. Method

2.1 Survey group

2.1.1 Courses and participants

A survey was conducted using two-unit 15-week Master's and Bachelor's courses, which were taught by the same professor at a Japanese national university. These two credit courses, which were organized as blended learning, were offered during the 2006 and 2007 Spring Terms. One course was "Information Society and Jobs", a Bachelor-level class for university freshman, and the other one was "Advanced Information Industry Studies", a Master's degree course for students in their first year of graduate studies. Most of the students were Engineering majors. The third course was "Information Society and Jobs", the same 2-unit Bachelor-level class for university freshmen, which was offered as a fully online course in 2006 and 2007. Students could choose to attend either the blended or the fully online course, in accordance with their preferences. Since freshmen in Japanese universities are busy attending various other courses, the fully online course can provide a flexible style of learning, if students are self motivated. Therefore, it is clear that fully online learning requires participants to possess a degree of information literacy and time management skills.

The total number of students with valid survey data was 184 (92 Masters; 67 Bachelors for blended learning, and 25 Bachelors for fully online learning).

2.1.2 Online courses

For the blended learning course, students attended face-to-face classes, and were also able to access the course content online outside of class. This online learning material was designed for a fully online course, and also for a blended learning course, to encourage students to maximize their learning experience. The materials, which were created in advance, consisted of lecture videos, slides and online tests which allowed students to check the progress of their learning achievement themselves. In particular, these online tests offered unlimited trial tests for students, and this system recorded the number of trials and the final test scores.

To encourage maximum participation in e-learning, in particular for the blended learning course, an explicit benefit was provided to students: online module test scores would count towards their final grades in the course. Also, students could make up for class absences by taking and passing online tests that corresponded with the face-to-face class sessions which were missed. This encouraged students to participate in online modules and tests, because if they missed a regular face-to-face class session it often affected their final test scores and the evaluation of their learning experience. Most students paid careful attention to their performance and final grades.

All students took part in the final test, which consisted two essay tests at the end of the course.
2.2 Survey instruments and data

All classes were surveyed using the same questions and constructs used in our earlier surveys (Nakayama et al. 2006). Several surveys of participants were conducted, and the results of these have already been reported (Nakayama et al. 2009). In this paper, we will focus on self-assessment of the online learning experience. To explain the differences in responses to questions, various indices were analyzed such as students’ information literacy and learning performance, and their essay writing ability.

2.2.1 Learning experience

The construct used to measure students’ online learning experiences consisted of a 10-item Likert-type questionnaire. This construct was originally developed by the authors to assess student’s activity during a blended learning course. It has been used repeatedly and its validity, including the factor structures has been confirmed over three years. The questions are shown in Table 2. The questionnaire, which asked about student’s overall impression of the online course, their own learning habits, and their learning strategies, required each item to be rated using a 5-point scale. The surveys were first conducted during the second week of classes and were then conducted again at the ends of the courses.

2.2.2 Information literacy

Information literacy as a characteristic of students was measured using a construct (Nakayama et al. 2008). Fujii (2007) has defined and developed inventories used to measure information literacy. For this construct, the survey consisted of 32 question items, and 8 factors were extracted: interest and motivation, fundamental operating ability, information collecting ability, mathematical thinking ability, information controlling ability, applied operating ability, attitude, and knowledge and understanding. This construct was surveyed during the second week of the courses.

Secondary factor analysis was conducted on the scores of the above eight factors which were calculated using the survey data. As a result, two secondary factors were extracted (Nakayama et al. 2008). The first secondary factor (IL-SF1) consists of “operational confidence and knowledge understanding”; the second one (IL-SF2) consists of “attitude issues”.

2.2.3 Learning performance

Some indices regarding learning performance during these courses were measured. Three indices were used as indicators of learning performance: the number of days attended (NDA), the number of completed modules (NCM), and the online test scores (OTS).

For Master’s students, the final test was conducted as a reporting-style essay. Master’s students wrote summary reports which were selected from two out of 5 topics. For Bachelors, the final test was conducted with a proctor during the final exam period assigned by the university. All Bachelor students gathered in a lecture room, and wrote answers to four questions. Two questions included multiple-choice tasks and the other two were essay tests.

The essay tests were evaluated by experts using an automated system (Nakayama et al. 2009). First, the essay tests were assessed by two outside experts. They independently evaluated essays using a 3-point scale (0-2) which used 5 criteria: certainty, fitness for learning content, argument, various aspects and drawing illustrations. The sum of 5 scores were used as an expert assessment. Second, for the assessment of the essay, an automated scoring system (Ishioka and Kameda, 2003) was used. It is possible to use this system via a web site. As a result, another score which was calculated using assessment software, consisted of three factors: "rhetoric", "logical structure" and "content fitness".

All data were analyzed to extract differences in survey timings using simple statistical tests, such as t-tests, correlation analysis and factor analysis (Coolican 1994).
3. Results

3.1 Learning activities

To compare the learning activities of Master’s and Bachelor students taking online courses, indices of learning, rates for the number of days attended (NDA), and rates for the number of completed modules (NCM) are summarized in Table 1. This table shows mean rates and STDs of indices, and the mean difference between Master’s and Bachelor students in the blended learning environment is tested statistically using a t-test, in order to make a clear comparison. The results and t-values are indicated in Table 1. There are significant differences in NCMs between Master’s and Bachelor students (p<0.01), while there are no significant differences in NDAs. According to our previous analysis (Nakayama et al., 2006), Master’s students prefer online modules more than Bachelors do, and this result confirms the tendency. Rates of days attended were quite high, so there was no difference in this measurement for Master’s and Bachelor students.

As a reference, the rate of the number of completed modules (NCM) for Bachelors in the fully online learning course is indicated in Table 1. The rate is comparable with the rate for the blended learning environment.

Table 1: Summary of learning activities

<table>
<thead>
<tr>
<th></th>
<th>Master (N=92)</th>
<th>Bachelor (N=67)</th>
<th>Bachelor(FO) (N=25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N of days attended (NDA)</td>
<td>0.95 (0.08)</td>
<td>0.95 (0.06)</td>
<td>t(157)=0.7 N/A</td>
</tr>
<tr>
<td>N of completed modules (NCM)</td>
<td>0.91 (0.17)</td>
<td>0.84 (0.15)</td>
<td>t(153)=2.7** 0.86 (0.10)</td>
</tr>
</tbody>
</table>

( ): STD, **: significant level p<0.01, FO: Fully Online course

As a reference, the rate of the number of completed modules (NCM) for Bachelors in the fully online learning course is indicated in Table 1. The rate is comparable with the rate for the blended learning environment.

Table 2: Question items and mean scores for Master’s and Bachelors (blended and fully online) students at the beginnings and the ends of courses

<table>
<thead>
<tr>
<th>Question items</th>
<th>Master(BL)</th>
<th>Bachelor(BL)</th>
<th>Bachelor(FO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. E-learning is easy to follow and understand</td>
<td>3.2 I 3.4 I</td>
<td>3.2 I 3.4 I</td>
<td>3.8 I 4.1 F</td>
</tr>
<tr>
<td>Q2. I learn better in an on-line course</td>
<td>2.3 &lt;&lt; 2.8</td>
<td>2.5 &lt; 2.9</td>
<td>3.3 I 3.2 I</td>
</tr>
<tr>
<td>Q3. On-line materials are useful to me</td>
<td>3.3 I 3.4</td>
<td>3.4 I 3.4</td>
<td>3.9 I 4.0</td>
</tr>
<tr>
<td>Q4. It is easy to schedule on-line learning time</td>
<td>3.5 I 3.7</td>
<td>3.2 I 3.2</td>
<td>4.2 I 3.9</td>
</tr>
<tr>
<td>Q5. On-line course content is interesting</td>
<td>3.2 I 3.2</td>
<td>3.2 &gt; 3.0</td>
<td>3.8 I 3.7</td>
</tr>
<tr>
<td>Q6. Overall, on-line course is a favorable learning experience</td>
<td>3.4 I 3.6</td>
<td>3.6 I 3.4</td>
<td>4.0 I 4.0</td>
</tr>
<tr>
<td>Q7. I'm a conscientious student</td>
<td>2.7 I 2.7</td>
<td>2.8 I 2.8</td>
<td>4.4 I 4.0</td>
</tr>
<tr>
<td>Q8. It is my habit to do course preparation and review</td>
<td>2.6 I 2.5</td>
<td>2.7 &gt; 2.5</td>
<td>2.7 I 2.5</td>
</tr>
<tr>
<td>Q9. I have my own method and way of learning</td>
<td>3.3 I 3.3</td>
<td>2.8 I 2.8</td>
<td>3.5 I 3.2</td>
</tr>
<tr>
<td>Q10. I have my own strategies on how to pass a course</td>
<td>3.7 I 3.7</td>
<td>2.7 &lt; 2.9</td>
<td>2.9 I 2.8</td>
</tr>
</tbody>
</table>

Significant level <<: p<0.01; <: p<0.05; I: Initial, F: Final, BL: Blended Learning, FO: Fully Online course

3.2 Mean values for question items in transition

Mean values for the 10 question items reveal some differences in means between the beginnings and the ends of the courses, and are summarized in Table 2. To determine the changes in responses, statistical tests such as t-tests of the initial and final mean scores have been conducted and the significant differences are marked with "<<" symbols according to the level of significance.

For example, for question item Q2 ("I learn better in online courses"), the means of both Master’s and Bachelor students in the blended learning environment increased significantly from the beginnings to the ends of the courses.

This result suggests that most participants prefer to learn using online materials along with the course. On the other hand, means for Q5 and Q8 decreased at Bachelor student levels in the blended learning course. This means that some problems have occurred with Bachelor students in the blended learning environment.
environment. A support system should be provided for these students because they are freshmen and do not have experience with the learning environment.

There was no significant difference in means between the beginning and the end of the fully online course, however most means of assessment were higher than the means of assessment for the blended learning course. Participants in the fully online course were satisfied with the learning setting and recognized the benefit of the course.

The results coincided with the results of our previous report (Nakayama et al. 2007), and the tendency for the responses to be the same was validated.

3.3 Factor score transitions

The model for the 10 survey questions, which had been used previously, consisted of three factors (Nakayama et al., 2009). Again the factor structure from the analysis of 456 participants across three years (Nakayama et al., 2009) was used. The contribution ratio and correlation across factor axes is summarized in Table 3. For this survey, the factor structure was used to determine students' attitudes toward temporal transitions.

Table 3: Correlations between factor scores

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. E-learning is easy to follow and understand</td>
<td>0.70</td>
<td></td>
<td>0.10</td>
</tr>
<tr>
<td>Q2. I learn better in an online course</td>
<td>0.48</td>
<td>0.20</td>
<td>-0.08</td>
</tr>
<tr>
<td>Q3. Online materials are useful to me</td>
<td>0.61</td>
<td>0.10</td>
<td>-0.10</td>
</tr>
<tr>
<td>Q4. It is easy to schedule online learning time</td>
<td>0.57</td>
<td>-0.16</td>
<td>0.18</td>
</tr>
<tr>
<td>Q5. Online course content is interesting</td>
<td>0.68</td>
<td>0.11</td>
<td>-0.05</td>
</tr>
<tr>
<td>Q6. Overall, online course is a favorable learning experience</td>
<td>0.82</td>
<td>-0.07</td>
<td>-0.02</td>
</tr>
<tr>
<td>Q7. I'm a conscientious student</td>
<td>-0.04</td>
<td>0.51</td>
<td>0.07</td>
</tr>
<tr>
<td>Q8. It is my habit to do learning preparation and review</td>
<td>0.11</td>
<td>0.49</td>
<td>0.08</td>
</tr>
<tr>
<td>Q9. I have my own method and way of learning</td>
<td>-0.01</td>
<td>0.18</td>
<td>0.62</td>
</tr>
<tr>
<td>Q10. I have my own strategies on how to pass a course</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.65</td>
</tr>
</tbody>
</table>

The factor scores in this survey are estimated using means of marked scale values for factor items. Here, the focus is on the initial and final factor scores. The factor scores for Master's and Bachelor students for blended and fully online courses at the beginnings and ends of courses are summarized in Figure 1. As the figure shows, there are no significant differences in mean scores between the initial and the final scores of courses using t-tests, while there is a significant difference in the third factor scores between Master's and Bachelor students for blended learning (Nakayama et al., 2007, 2009). This result confirms the coherence of participants' evaluations of their learning experiences for both blended and fully online courses.

The course was conducted across 15 weeks, and participants felt strange about recording their impressions of their progress during the course. A detailed analyses was conducted as follows. According to the procedure for factor analysis, factor scores correlate with each other partially because the structure of the three factors is extracted as a Promax solution (Nakayama et al., 2009). The correlational relationships in this survey were confirmed, as shown in Table 3. This result suggests that the second factor scores correlate with both the first and third factor scores.
3.4 Relationships of factor scores between the beginnings and the ends of courses

To determine the coherence of factor scores during a course, factor scores of the relationships were analyzed. Figure 2 shows a scatter gram, which illustrates pairs of the first factor scores between the beginnings and ends of the Master’s and Bachelor (blended and fully online) courses. The regression lines were superimposed over the figures for Masters and Bachelors blended and fully online courses respectively. There were some differences in slopes of the three regression lines.

These superimposed figures suggest that factor scores between the beginnings and ends of courses do not coincide. Therefore, individual factor scores deviate during the courses.

On the other hand, the relationships between factor scores for Factor 2 is illustrated in Figure 3. Since scores of Factor 2 are calculated from responses to two items, the plots are sparse. Most plots are gathered along a diagonal line, and both scores are correlated, though some deviations can be seen in the figure. The regression lines almost overlap each other without much difference in the angles of the slopes.

---

**Figure 1:** Comparison of factor scores between Master’s and Bachelors (blended and fully online) students at the beginnings and the ends of courses.

**Figure 2:** Scatter gram of factor 1 scores between the beginnings and the ends of courses.
To confirm this tendency, correlation coefficients of factor scores between the beginnings and the ends of the courses are summarized in Table 4 for Master’s students, in Table 5 for Bachelor students in the blended course and in Table 6 for Bachelor students in the fully online course. The vertical cells show the factor scores at the beginning (I: initial), and the horizontal cells show the factor scores at the end (F: final). For the three tables, diagonal coefficients show the degree of deviation of the factor scores. As all coefficients between the second factor scores are above 0.7, this assessment aspect may be deemed to be stable. For the third factor score, the coefficient for Bachelor students in the blended course (r=0.37) is relatively small because many freshman Bachelors have acquired new learning strategies during the course, though a number of Bachelors who had already become information literacy experts participated in the fully online course instead of the blended course. For the first factor score in all Master’s and Bachelor classes, the coefficients are relatively small (Master: r=0.35, Bachelors in blended: r=0.46, Bachelors in fully online: r=0.33), so student’s assessments may have changed during the courses.

**Table 4: Correlation coefficients of factor scores between the beginnings (I) and the ends (F) of courses (Master’s)**

<table>
<thead>
<tr>
<th>Master (Blended)</th>
<th>F1(F)</th>
<th>F2(F)</th>
<th>F3(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1: e-Learning overall evaluation(I)</td>
<td>0.35</td>
<td>0.24</td>
<td>0.15</td>
</tr>
<tr>
<td>F2: Learning habits(I)</td>
<td>0.20</td>
<td>0.72</td>
<td>0.34</td>
</tr>
<tr>
<td>F3: Learning strategies(I)</td>
<td>-0.01</td>
<td>0.32</td>
<td>0.51</td>
</tr>
</tbody>
</table>

**Table 5: Correlation coefficients of factor scores between the beginnings (I) and the ends (F) of courses (Bachelor: Blended)**

<table>
<thead>
<tr>
<th>Bachelors(Blended)</th>
<th>F1(F)</th>
<th>F2(F)</th>
<th>F3(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1: e-Learning overall evaluation(I)</td>
<td>0.46</td>
<td>0.22</td>
<td>0.15</td>
</tr>
<tr>
<td>F2: Learning habits(I)</td>
<td>0.43</td>
<td>0.70</td>
<td>0.46</td>
</tr>
<tr>
<td>F3: Learning strategies(I)</td>
<td>0.25</td>
<td>0.33</td>
<td>0.37</td>
</tr>
</tbody>
</table>

**Figure 3:** Scatter gram of factor 2 scores between the beginnings and the ends of courses.
Table 6: Correlation coefficients of factor scores between the beginnings (I) and the ends (F) of courses (Bachelor: Fully online)

<table>
<thead>
<tr>
<th>Bachelor(Full Online)</th>
<th>F1(F)</th>
<th>F2(F)</th>
<th>F3(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1: e-Learning overall evaluation(I)</td>
<td>0.33</td>
<td>0.34</td>
<td>0.30</td>
</tr>
<tr>
<td>F2: Learning habits(I)</td>
<td>0.25</td>
<td>0.71</td>
<td>0.45</td>
</tr>
<tr>
<td>F3: Learning strategies(I)</td>
<td>0.44</td>
<td>0.61</td>
<td>0.55</td>
</tr>
</tbody>
</table>

When observing correlation coefficients across the three factor scores, the coefficients between the first and the third factor scores are small for both Master’s and Bachelors students in the blended course but not for the fully online course. As shown in Table 3, the original factor scores are weakly correlated as well. In Tables 4, 5 and 6, the initial second factor scores correlate significantly with all final factor scores except for the first factor in the fully online course. The initial second factor scores show the extent of participant’s learning habits, and all factor scores increase during the courses when participants have acquired some additional learning habits, though the second factor axis correlates with the axes of the other two factors. Also, students whose learning habits are poor should be encouraged to improve them and be supported in doing this.

3.5 Students’ transitions during courses

According to the above analyses, assessment transitions may provide some critical information when individual differences in factor scores are extracted and correlation coefficients are calculated between the differences in factor scores at the beginnings and ends of courses. These coefficients are summarized in Table 7.

<table>
<thead>
<tr>
<th>Score difference</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial: Master(BL)</td>
<td>-0.60</td>
<td>-0.47</td>
<td>-0.54</td>
</tr>
<tr>
<td>Initial: Bachelors(BL)</td>
<td>-0.37</td>
<td>-0.34</td>
<td>-0.60</td>
</tr>
<tr>
<td>Initial: Bachelors(FO)</td>
<td>-0.65</td>
<td>-0.21</td>
<td>-0.44</td>
</tr>
<tr>
<td>Final: Master(BL)</td>
<td>0.54</td>
<td>0.28</td>
<td>0.44</td>
</tr>
<tr>
<td>Final: Bachelors(BL)</td>
<td>0.65</td>
<td>0.43</td>
<td>0.52</td>
</tr>
<tr>
<td>Final: Bachelors(FO)</td>
<td>0.46</td>
<td>0.61</td>
<td>0.49</td>
</tr>
</tbody>
</table>

All coefficients are significant: p<0.05 except where underlined

The coefficients between the differences and the final factor scores are also calculated, and all coefficients are positive and almost always high because when the final factor scores are high, the differences are large. Therefore, high factor scores come from final scores which have increased since the initial scores.

When survey assessments of the online learning environment are conducted, the timing of the survey may affect the results. Some participants become discouraged during the course, but the assessor will still receive a good assessment at the end of the course if evaluations come from satisfied students whose impressions have become more positive during the course. If the timing of the survey is at an earlier stage in the course, the structure of the responses may be the opposite.
3.6 Relationship between students’ transitions and other metrics

To examine the relationship between students’ transitions and other metrics, participants who have all metrics were selected. Therefore, the number of students decreased to 78 Master’s and 45 Bachelors students for blended learning, and 25 Bachelors for the fully online course.

First, the impact of student’s information literacy (IL) on students’ transitions and on the assessment factor scores is confirmed. For Master’s students in the blended learning environment, both information literacy, “operational confidence and knowledge understanding” (IL-SF1) and “attitude issues” (IL-SF2) affect both initial and final assessment factor scores. There was no relationship between information literacy and differences in factor scores though some correlational coefficients changed between the initial and the final surveys. Some negative coefficients appeared between the first factor (F1) and learning performance, such as NDA and OTS, because the indices of learning performance of participants decreased in accordance with degree of their preference for online learning (F1). There were no significant relationships for other factors. Also, performance in the essay test did not affect the factor scores of students’ transitions.

For Bachelor students in the blended learning environment, the first information literacy measure: “operational confidence and knowledge understanding” (IL-SF1) affected factor scores of students’ transitions. There are some positive coefficients between learning performance and students’ transitions contrary to Master’s students. Most positive coefficients suggest that scores for “learning habits” correlate with the number of days (NDA) of face-to-face sessions which were attended, and scores of “learning strategies” correlate with the number of completed online modules (NCM). The difference in the first factor positively correlates with the NCM (r=0.32), since these participants have recognized the benefits of online learning. Positive evaluation of online learning (F1) may influence essay scores, experts assessment scores (EXP) and automated evaluation scores (Auto scores), because the coefficients are negative.

<table>
<thead>
<tr>
<th>Masters (BL)</th>
<th>Initial F1 F2 F3</th>
<th>Final F1 F2 F3</th>
<th>Difference F1 F2 F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-SF1</td>
<td>0.22 0.40 0.27</td>
<td>0.01 0.28 0.17</td>
<td>-0.23 -0.15 -0.16</td>
</tr>
<tr>
<td>IL-SF2</td>
<td>0.23 0.48 0.23</td>
<td>0.20 0.49 0.29</td>
<td>-0.04 0.03 0.00</td>
</tr>
<tr>
<td>NDA</td>
<td>-0.17 -0.09 0.03</td>
<td>-0.38 -0.09 -0.03</td>
<td>-0.25 0.00 -0.06</td>
</tr>
<tr>
<td>NCM</td>
<td>-0.18 0.07 -0.04</td>
<td>-0.19 0.01 0.03</td>
<td>-0.02 -0.08 0.09</td>
</tr>
<tr>
<td>OTS</td>
<td>-0.02 -0.02 0.08</td>
<td>-0.36 -0.08 0.05</td>
<td>-0.37 -0.08 -0.04</td>
</tr>
<tr>
<td>EXP score</td>
<td>0.04 -0.02 -0.03</td>
<td>-0.04 -0.01 -0.11</td>
<td>-0.09 0.01 -0.06</td>
</tr>
<tr>
<td>Auto score</td>
<td>0.10 -0.15 0.09</td>
<td>0.01 0.04 0.22</td>
<td>-0.10 0.08 0.10</td>
</tr>
</tbody>
</table>

IL-SF1: operational confidence and knowledge understanding, IL-SF2: attitude issue
NDA: N of days attended, NCM: N of completed modules, OTS: Online test score
EXP: Expert essay assessment score, Auto: Automated essay assessment score

For Bachelors students in the fully online course, there are a few positive coefficients between information literacy and factor scores of students’ transitions. Because the differences in factor scores between the beginning and the end is small and the number of valid participants is not large, the number of significant correlational coefficients is small. During online courses, learning activities may not be affected because there are no face-to-face sessions and no collaborative learning activities amongst fellow students. According to this hypothesis, students’ transitions may be due to collaborative learning activities where instructors and fellow students have collaborated during face-to-face sessions.

The confirmation of this hypothesis will be a subject of our further study.
### Table 9: Correlational relationship between transitions of factor scores and learning performance (Bachelors: Blended learning)

<table>
<thead>
<tr>
<th>Bachelor (BL)</th>
<th>Initial</th>
<th>Final</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>F2</td>
<td>F3</td>
</tr>
<tr>
<td>IL-SF1</td>
<td>0.35</td>
<td>0.21</td>
<td>0.26</td>
</tr>
<tr>
<td>IL-SF2</td>
<td>0.15</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>NDA</td>
<td>-0.06</td>
<td>-0.13</td>
<td>0.32</td>
</tr>
<tr>
<td>NCM</td>
<td>0.18</td>
<td>0.07</td>
<td>0.35</td>
</tr>
<tr>
<td>OTS</td>
<td>-0.32</td>
<td>-0.20</td>
<td>-0.25</td>
</tr>
<tr>
<td>EXP score</td>
<td>0.10</td>
<td>-0.23</td>
<td>-0.19</td>
</tr>
<tr>
<td>Auto score</td>
<td>0.07</td>
<td>-0.23</td>
<td>-0.19</td>
</tr>
</tbody>
</table>

IL-SF1: operational confidence and knowledge understanding, IL-SF2: attitude issue, NDA: N of days attended, NCM: N of completed modules, OTS: Online test score, EXP: Expert essay assessment score, Auto: Automated essay assessment score

### Table 10: Correlational relationship between transitions of factor scores and learning performance (Bachelors: Fully online learning)

<table>
<thead>
<tr>
<th>Bachelor (FO)</th>
<th>Initial</th>
<th>Final</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>F2</td>
<td>F3</td>
</tr>
<tr>
<td>IL-SF1</td>
<td>0.11</td>
<td>0.43</td>
<td>0.34</td>
</tr>
<tr>
<td>IL-SF2</td>
<td>-0.03</td>
<td>0.37</td>
<td>0.29</td>
</tr>
<tr>
<td>NDA</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NCM</td>
<td>0.11</td>
<td>-0.35</td>
<td>-0.16</td>
</tr>
<tr>
<td>OTS</td>
<td>-0.14</td>
<td>0.31</td>
<td>0.17</td>
</tr>
<tr>
<td>EXP score</td>
<td>0.25</td>
<td>0.15</td>
<td>0.38</td>
</tr>
<tr>
<td>Auto score</td>
<td>0.07</td>
<td>0.12</td>
<td>0.23</td>
</tr>
</tbody>
</table>

IL-SF1: operational confidence and knowledge understanding, IL-SF2: attitude issue, NDA: N of days attended, NCM: N of completed modules, OTS: Online test score, EXP: Expert essay assessment score, Auto: Automated essay assessment score

### 4. Conclusion

In this paper, assessments of participants' transitions were analyzed to determine the effectiveness of online learning which included blended learning and fully online courses. The mean factor scores, which were extracted as a measure of learning experience, remained at the same levels between the beginnings and the ends of the courses. The individual assessment scores for e-learning evaluations changed dramatically, though the scores for learning habits were relatively stable. The differences in factor scores during the courses correlate positively with the final scores, while these differences correlate negatively with the initial scores. Additionally, several indices of learning performance were surveyed, and there were some relationships between a number of these indices and participants' transitions in the blended learning course, while there were few such relationships in the fully online learning course.

Therefore, the course assessors should bear in mind the timing of the survey and the significance of the results. The process of transition during the course should be analyzed more closely. Also, some effective support procedures for participants need to be developed. A detailed discussion of these items will be the subjects of our further study.

### Acknowledgments

This research was partially supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (B19300274: 2007-2009).

### References


Science, Sport and Technology - a Contribution to Educational Challenges

Kelly O'Hara¹, Paula Reis², Dulce Esteves¹, Rui Brás¹ and Luísa Branco¹
¹FCSH, University of Beira Interior, Covilhã, Portugal
²Information, Verification, Service, USA

Abstract: Improve students' ability to link knowledge with real life practice, through enhancing children or teenagers' ability to think critically by way of making observations, posing questions, drawing up hypotheses, planning and carrying out investigations, analysing data and therefore improve their decision making is an educational challenge. Learning through sports can be effective for developing life skills because sport has a potential to contribute over a wide range and is a discipline that most children like. The constructions of real situations or "Problems" must achieve and incorporate certain aspects such as (a) encourage curiosity, (b) be perceived by students as relevant to their personal goals, (c) represent a motivated challenge, (d) stimulate group collaboration for older students, (e) technological equipment as a way of support, to motivate the learning process, and (f) demonstrate how simple scientific concepts can improve everyday activities. The aim of this paper is to present and evaluate the usefulness of the representative tasks created by a systematic integration of approaches (electronic and non-electronic devices) with interactive situations. Four tasks were applied to 140 children between 6-10 years old at elementary school level. The tasks were constructed considering the follow proposals: (1) promote the benefit of physical activity and (2) explore some science concepts using sport. To evaluate the process effectiveness, two groups were formed, group A was submitted to a more theoretical explanation of the concepts and group B was exposed to problem solving through sport situations. Data were analysed by using quantitative methods. Results show that when children participate in an active way they are more motivated, and the use of their own movement or body to resolve a problem (with electronic devices) contributes for knowledge acquisition by adapting their actions and looking for the best window of possibilities to solve the task situation. Further and longitudinal studies are recommended to consolidate the results.

Keywords: technology, sport, task design, skills acquisition

1. Introduction

The needs and demands of education in the modern dynamic world has become one of the greatest challenges we now face (Dewey1997a, b; Delacôte, 1998a, Fishman & McCarthy,1998). No longer is one teacher, one classroom, one desk and one textbook sufficient. We also know that it is not merely a question of resources, finances, material or technology, items which although important, will not in themselves solve the number one problem, which is how to supply a wider range of competences for students to be well prepared for a future which is indelibly shaped by uncertainty (Colwell &Kelly, 1999; Windschitl, 2000; Wilson & Corbett, 2000). ELearning implementation enables the learner to explore, discover, and construct subject matters and ideas with other community members, being a powerful tool to achieve goals if teachers have suitable strategies and pedagogies for an effective implementation (Kahiigi, Ekenberg, Hanssson, Tusubira & Danielson, 2008; Garrison, Anderson & Archer, 2000).

When we observe the educative landscape at the macro level, we can verify (OECD, 2002) that for some, we are in (a) an Information Society, while for others, in (b) a Knowledge Society, and for yet others, in (c) a Learning Society. However, as we concentrate on supplying a wider range of competences, it is absolutely crucial to analyse the consequences and implications of being in these "societies", since:

(a). to be in the Information Society is to run the risk that the information received is just as likely to be disinformation, because the quantity of data available nowadays (which may or may not be useful, correct or incorrect, or may have a positive or negative impact), whether in terms of quantity or speed, is only worthwhile if it allows the user to develop the necessary competences that permit the user to
select from the available information, that which will contribute to a better understanding, solve problems and learn new things;

(b). to be in the Knowledge Society implies understanding the difference between knowledge as an unrelated set of facts and knowledge as an integrated phenomenon and also understanding if that which is valued most, is or is not how the individuals obtained that knowledge, the things they picked up on the way to getting that knowledge and what motivated the learning of that knowledge;

(c). to be in the Learning Society where students have teachers and other resources at their disposal to achieve their goals and objectives, implies having the capacity to prepare competent individuals who will want and know how to implement projects, because, according to the OECD, a society that demands an ASK programme (attitudes, skills, knowledge).

If we zoom inside this educational landscape at the micro level, we need to have the capacity to develop the necessary competences which consider learning to be largely about modifying individual representations rather than the result of a process of “stacking” information/knowledge. A representation is not just the outcome i.e. that which the student verbalizes, writes, draws, plays or does, but the underlying neuronal structure from which these actions originate (Giordan, 1998).

Guided by the above mentioned macro and micro perspectives, we propose in this paper to design an experimental approach to work on an example of this wider range of competencies, Problem-Solving Skills, “i.e. the capacity of students to understand problems situated in novel and cross-curricular settings, to identify relevant information or constraints, to represent possible alternatives or solution paths, to develop solution strategies, and to solve problems and communicate the solutions” (OECD, 2004), showing both the theoretical foundations that need to be considered, and the active principles that we turn to for operational purposes.

From a theoretical point of view, the following are to be considered:

Learning should be centred on the student

In order to construct an interactive environment as a tool to assist students in their reasoning and finding the possible solutions, they become active participants in the learning process. However even when centred on the student, this process needs to be structured and facilitated by the teacher and/or team work, in order to permit the realization of concrete situations, where the activities being promoted will allow the participants to manipulate, experiment and interpret phenomena according to different functional perspectives and previously defined objectives (Dewey1986, 1997; OECD, 2002, 2004, 2007);

The development of a process to stimulate curiosity based on problem situations

The capacity to raise pertinent questions and attempt to critically interpret the available indicators in order to find the best possible solutions depends on a change of attitude and the profound transformation of the individual, while demanding systematic and sustainable work. Clearly this work rests on the capacity of the team work to make use of conceptual and material instruments, while trying to establish a ludic and critical spirit in the search for organisation and mastering of knowledge (Richard, 1999; OECD, 2002).

The development of a technological and scientific reasoning

The problem situations put forward by the team work should stimulate curiosity and the capacity to be critical which in turn will lead to the development and training of a scientific reasoning that will a) raise and equate hypotheses for resolution; b) use information, data, and available documentation to allow reflection, and represent systems for alternative resolutions; c) select strategies for resolution; d) present results and reasons which can be debated; d) present new questions and doubts which appear as the process runs its course (Folder, 1993, Alper, 1994; Delacôte, 1998b; Richard, 1999; Crawford, Saul, Mathews & Makinster, 2005; Reimann & Jacobson, 2010).

On the bases of the above mentioned theoretical foundations, our experimental situation (Delisle, 1997) made use of the following active principles:(
Student-centred learning

Our interactive environment is done through real life situations which, depending on the students' level, will allow them to create strategies to solve the problems. What is desired here is that when students are presented with a series of problems, they will solve them by making the best choice from a window of possibilities. To accomplish this we used various instruments, including computer graphics, animations, heart rate monitors, videos and smart boards.

Developing a process of curiosity based on problem situations.

Sport is used as the catalysing agent for the motivation and interest of the participants: the use of games to present strategic problems guided by objectives and rules; the definition of competitive situations within a certain range of challenge; the elaboration of a points system that will permit the verification of results and the fulfilment of partial objectives, a system for rotating through different goals and or tasks, allows the acknowledgement of individual limits and identify the strong and weak points of each element leading to the improvement of the team work while at the same time showing how different individual competences will affect the capacity of the team.

The development of a technological and scientific reasoning.

The tasks were conceived for promoting the benefit of physical activity and to explore some science concepts. All situations considerer different resolutions and strategies to reach the goal and are accompanied with video and smart board information. This information was given: a) at the beginning of the task, so we can present the problem, the concept and the aiming goal, b) during the task for helping the possible problem resolution and c) at the end in order to make an overview of the global task and provoking discussion about the concepts and its functionality.

In this article we chose the concepts of heart rate (HR) and energetic balance (EB). Physical activity and health are recognized as an important means to helping children and youths attain a healthy emotional, social and physical well being (Strong, Malina, Bumkie, Daniels, Dishman, Gutin, Hergenroeder, Must, Nixo, Pivanirk, Rowland, Trost & Trudeau, 2005). In fact, encouraging the physical activity in children and adolescents is often viewed as an effective health promotion and a disease prevention strategy for the adult population (Thompson, Humbert & Mirwald, 2003). In last decades numerous studies reported a significant decline in physical activity levels on children and adolescents. Different factors contribute for this physical inactive behaviour, typical of the actual lifestyle. The new technologies, sedentary player games, and the unsafe and inaccessible places to practice physical activity are some examples that do not support the regular physical activity prevalent in childhood and youth (Larsen, McMurray & Popkin 2000).

To promote a lifetime inclusion of physical activity in children and adolescents, several aspects should be considered for a globally agreed strategy, as school and physical education, community and sports local programs, and family support (Thompson et al., 2003). The school has an important role in this field, because it can promote the practice of diverse physical activities and experiences, but it can also promote the physical activity as an instrument of learning and social and physical development (Folder, 1993; Alper, 1994, Laursen, Liston, Thiry & Graf, 2007).

2. Methodology

In order to investigate the effectiveness of the active principles defined in micro level (points 1, 2 and 3), we create interactive real life tasks by integrating the concepts of Energetic Balance (EB) and Heart Rate (HR) with different eLearning tools (Kahiigi et al, 2008), such as video camera, heart rate monitor and smart board.

2.1 Participants

140 children (age between 6 to 10 years) from elementary school were selected and divided randomly in two groups, where group A, n= 73 and group B, n=67. Group A were exposed to a more theoretical explanation for heart rate (HR) concept but made the practical situations for energetic balance (EB) whereas group B made the opposite approach.
2.2 Statistical analysis

Descriptive analyze and independent samples Mann-Whitney U test was used to compare groups. A $p$ value of 0.05 or less was considered as significant. All $p$ values are two-tailed. The data were analyzed using the software Statistical Package for the Social Sciences (PASW 18 for Windows; SPSS).

2.3 Experimental design

All tasks (theoretical and practical) started with a question based on simple real things and built in a way that children were conducted to think, to encourage curiosity, to motivate solving the problem and the answer could only be achieved if they played the game. The introductory purpose for each task gives the clue to the concept that will be introduced. The purpose is complete enough to present the goal, but does not give away the “mystery” of the results. There is an element of surprise in each tasks and guaranteed fun and competition.

For the EB the questions were “can we have the same energy from all types of food?” and “can we eat without being fat?”. To the HR concept the questions were “How does our blood circulate?” and “how can we increase our heart rate?”

At the end of each task, a test for each concept was applied, equal for both tasks (theoretical and practical). More than evaluating the capacity of children to define the concepts per si, we intend to assess the capacity of the learning tasks proposed to develop the aspects presented on micro level (points 1,2 and 3).

For EB concept the test has a total of 30 questions, divided in 3 main categories, where Q1 and Q2 correspond to the capacity of the children to define the concept (with a total of 12 questions, 6 for Q1 and 6 for Q2), Q3 and Q4 represents the capacity of the children to choose and combine different possibilities (a total of 10 questions, 6 for Q3 and 4 for Q4) and Q5 and Q6 the capacity that they have to integrate different aspects of the learning knowledge to give the answer (a total of 8 questions, 5 for Q5 and 4 for Q6).

For HR concept the test has a total of 32 questions. The same structure was used to define the categories, where Q1 and Q2 has a total of 11 questions, 5 for Q1 and 6 for Q2, Q3 and Q4 have a total of 12 questions, 6 for Q3 and 6 for Q4 and Q5, Q6 and Q7 have a total of 9 questions, 5 for Q5, 2 for Q6 and 2 for Q7.

2.4 Theoretical tasks

The theoretical task (both for EB and HR) started with a question about the concept definitions, their significance, how they can be quantified and altered. Visual and experimental support instrumentation was used to facilitate the concept understanding. Questions were presented in order to create interaction between children and research team for establishing the comprehension level of the concepts by them. In consequence, the research team was able to apply different types of games.

For the HR concept the questions were “How does our blood circulate?” and “how can we increase our heart rate?”. The tasks were created in order to develop the perception of the heart rate concept, monitoring with polares (HR instrument).

For the concept of EB the introductory questions were “Can we have the same energy from all types of food?” and “Can we eat without being fat?”. The tasks aimed to provoke the sensibility of the concepts of calorie and fat cell. During the task the children had video segments and smart board information, for instance, they calculated how much energy was introduced to the organism by some food (visible on video) and how much was used (for example, in relation to sports activity).

Research team used those tasks not only to diagnose the comprehension level but also to control the children’s evolution.

2.5 Practical tasks

With the intention of establishing the level of children’s knowledge, in the beginning of the task, they were questioned about day-to-day things where the concepts (HR and EB) could be applied. At the
end of each task, they were once more questioned not only about the definition of the concept but also about the underlying aspects of the concepts and its functionality, such as “What did you do to reach your objective?”, “How did you do it?”, “Why did you do this?”.  

For the concept of Heart Rate (HR) the main question was “How does our blood circulate?” and “how can we increase our heart rate?”. Children show that they have the perception that only physical effort can increase HR. Various type of games were used to show different levels of HR increase and the recovery time needed to decrease those different intensity levels. The children had their own HR monitor as a feedback for their performance.

Since HR variations also depend of other factors, two tasks were set up to show how emotions can influence HR:

Heart rate task 1 (HR1) - “Let’s Dance”, the aim is to show that stress resulting from the relationship between two people could increase HR. Research team manipulated territorial behavioral space and children were paired up, male and female. Each child had a HR monitor beeper according to the maximum and minimum levels set by the supervisor. The goal was to dance without passing the HR maximum limit. If this happens they had different possibilities to decrease the HR value and return to the game. The points were awarded based on their capacity to control HR. At the end, by using video images, we could show them how stress led to an increase in HR.

Heart rate task 2 (HR2) - “Treasure hunting”, the goal is to show that factors like blindfolded and communications between the pair provoke the HR changes. The children were put in pairs with the objective of seeing which pair could find the largest number of balls (the treasure). In each pair, one child was blindfolded and could only move by following his partner’s verbal instructions. The children won the game if they could catch more balls, without passing the HR maximum limit.

More in both HR situations the player had different possibilities, by managing their own action and/or the opponent team, to increase or decrease HR values.

For the concept of Energetic Balance, the main questions were “Can we have the same energy from all types of food?” and “Can we eat without being fat?”. Two tasks were set up:

Energetic Balance situation 1 (EB1) - “Eating without being fat”. The aim was to show that calories are fundamental in supplying energy to the organism; if they are not completely used up they contribute to weight increase. Each pair received a card with a food type. First they had to identify the quantity of calories of this food type (using video information) and then define the effort that they have to make to burn the excess calories, in proportion to the amount of effort needed to complete the task. They were then asked to complete a circuit with more or less difficulties (by the type of obstacles, travel distance), depending on the number of calories per effort.

Energetic Balance situation 2 (EB2) - “Looking for the good fat cell”. The aim was to show that there are various types of fat cells (good and bad) and their consequent implications on the functioning of the organism. To transmit this knowledge researches create a circuit that combine speed, travel distance, balance and skill accuracy. To transmit the notion of bad fat cell they perform the circuit was mobility was decreased by adding weight in child limbs and by using uncomfortable clothes. For the notion of good cells they played normally without any constrains. The games began with each pair receiving a card with a food type and then perform the circuit.

3. Results and discussion

Heart Rate Tasks

Group B who perform practical situations present more correct answers than group A. Table 1 show the significative difference for both groups and the respectively $p$ value when applying Mann-Whitney U test for $p<0.05$.

In questions Q1 and Q2 (Figure 1 and 2) even with significative difference between groups for 2b(83% for PG and 54% for TG), the response tendency was the same (correct) but not in 2e (21% for PG and 50% for TG). One possible explanation is that 2e spotlighted on questioning the normal values of Hr for a healthy grownup where 2b focus on using equipment to measure HR. Further, it is
possible to observe that both groups give wrong answers when testing their concept knowledge about the phenomenon (Q1). One reason for this result can be the duration of the class thirty minutes plus the test, the age group in relation to the concept abstraction. In practical situations the children were so excited with the tasks environment that had difficulty to concentrate in the beginning of the class.

**Table 1:** p value for the answers in questions groups, when comparing PG Group B and TG Group A, p<0.05 Mann-Whitney U test

<table>
<thead>
<tr>
<th>Question Nº</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2b</td>
<td>.015</td>
</tr>
<tr>
<td>2e</td>
<td>.009</td>
</tr>
<tr>
<td>3a</td>
<td>.025</td>
</tr>
<tr>
<td>3f</td>
<td>.011</td>
</tr>
<tr>
<td>4b</td>
<td>.008</td>
</tr>
<tr>
<td>4f</td>
<td>.037</td>
</tr>
<tr>
<td>6b</td>
<td>.043</td>
</tr>
</tbody>
</table>

**Figure 1:** Percentage (%) of correct answers in Q1 for TG and PG in HR tasks.
Figure 2: Percentage (%) of correct answers in Q2 for TG and PG in HR tasks.

Although these questions present more variation between groups, the tendency of the answer is the same so the high values of the correct ones.

Figure 3: Percentage (%) of correct answers in Q3 for TG and PG in HR tasks.

Figure 4: Percentage (%) of correct answers in Q4 for TG and PG in HR tasks.
For wrong answers the values are similar in both groups, therefore no significative difference was observed when the children respond to question that they have to choose and combine different possibilities.

The same tendency is present for Q5, Q6 and Q7 only differing on the number of correct answers (table 2). The children were able to correlate different information (questions 6 and 7) independent of the way the knowledge was transmitted. Nevertheless in both tasks the teacher had the same conceptual strategies.

**Table 2: Percentage (%) of correct answers in Q5, Q6 and Q7 for TG (A) and PG (B) in HR tasks**

<table>
<thead>
<tr>
<th>Group</th>
<th>Question 5 (Q5)</th>
<th>Question 6 (Q6)</th>
<th>Question 7 (Q7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5a</td>
<td>5b</td>
<td>5c</td>
</tr>
<tr>
<td>PG</td>
<td>79</td>
<td>83</td>
<td>79</td>
</tr>
<tr>
<td>TG</td>
<td>71</td>
<td>88</td>
<td>71</td>
</tr>
</tbody>
</table>

Even with bad results in Q1, e.g., defining the concept, the children were able to correlate and choose the best answer when doing the comprehension question, according to the results showed in Table 3 and 4. These results strengthen the hands-on pedagogical process (and because even in the theoretical situations they had some experience by using e learning approach) and the problem based learning (Dochy, Segers, Van den Bossche & Gijbels, 2003; Wilson & Corbett, 2001).

Energetic Balance Tasks

Only for one question in group Q1 and Q6 there were significative difference, p=.003 and p =.001 respectively for p<.05 Mann-Whitney U test.

The results show that the percentage of correct answer are similarly (figure 5 and 6) with in groups (TG and PG) for the more memory questions.

**Figure 5:** Percentage (%) of correct answers between TG (B) and PG (A) in Q1 for EB tasks

Even when the children give the wrong answer (1d) both groups present a very low percentage. Therefore it is possible to see the same tendency of answer (correct or wrong) between groups.

In respect to those questions that involved the selection of more than one possibility Q3 and Q4 there were no difference between groups. Question 3a (figure 7) and 3 e present a significative difference between groups with a divergent tendency on their answers.
For the last questions groups (Q5 and Q6), despite only one question showing significant difference between groups, in table 7 it is possible to see that in all Q6 there is a great discrepancy in the answer given by the children. For TG with the exception of 5.3 and 5.5 (where they show that 11% and 33% give partially correct answer) wrong answers were given in this category. The PG children were able to provide correct answers whenever they had to make a relationship of the knowledge learned.

Table 3: Percentage (%) of correct answers between TG(B) and PG(A) in Q5 and Q6 for EB tasks

For a total of 62 questions only nine of them present significative difference between Practical (PG) and Theoretical (TG) groups (7 in 32 for HR and 2 in 30 for BI). These difference were in all questions groups (Q1 to Q7) but with more relevance in the group that had practical situations. Data presented in this paper is in agreement with Barrows (2004), OECD (2004) and Savin-Baden (2003), Laursen, et. al (2007) that children will benefit when exposed to a learning process constructed with the
premises on developing the ability to make decisions and solve problems, and when exposed to several bodies of knowledge in a real life context. Because most of the situations presented in research create the task using one kind of knowledge, more studies with the same framework behind the task constructions are needed, in other to support our results.

4. Conclusion and future works

The findings obtained from this study are in agreement with Windschitl (2002) that learning would be optimized if students were involved in real life and meaningful problem-based activities. Also the results show that when children were faced by practical situations they develop the capacity to understand different constrains, and adapt their actions looking for the best window of possibilities to solve the problem, since they score more correct answers that those children exposed just too theoretical situations. We can verify that the use of e learning methods integrated within the learning process, if constructed by the guide premises (macro and micro level) considering the realness and the relatedness of the problem to the children’s everyday experiences, can create a learning intention (Laursen, et al., 2007; Reimann & Jacobson, 2010). Participating in sport activity in an intentional way can develop life skills (Danish & Nellen, 1997, Goudas, Danish & Theodorakis, 2005), i.e. skills that are required to deal with the demands and challenges of everyday life. They can be not only physical (e.g., taking the right posture), but behavioral (e.g., communicating effectively), cognitive (e.g., making effective decisions). Nevertheless more studies are required, with the same conceptual structure, for developing new tasks with other concepts, or the same concepts with more variability and versatility of games and eLearning tools, increasing the integration of areas of knowledge and research teams. These new tasks should be applied to different ages and social contexts. Also, it is important to develop new assessment approaches and tools in order to control the reasoning process we present at micro level (1, 2 and 3). Longitudinal studies are also recommended to consolidate the results. The proposed tasks contribute to the development of the new educational challenge, showing how it can be possible to learn not only basic knowledge but also develop others skills and train the capacity to make decisions and solve problems in a motivating environment. Another conclusion from this study is that technology support and eLearning processes can be a powerful tool for learning when integrated within a conceptual framework that considers the development of Problem-Solving Skills, the ability to think critically and therefore improve decision making. In this case technology will naturally be included. So creating tasks that integrate different kind of knowledge, technology and in a same way are pleasure for does who participate, benefits the educational process and facilities the teacher action.

References


www.ejel.org 96 ©Academic Publishing International Ltd


From Soap Opera to Research Methods Teaching: Developing an Interactive Website/DVD to Teach Research in Health and Social Care

Abigail Sabey and Sue Horrocks
University of the West of England, Bristol, UK
Abby.Sabey@uwe.ac.uk
Susan.Horrocks@uwe.ac.uk

Abstract: Research methods modules have become a core component of a range of nursing and allied health professional educational programmes both at pre-qualifying, undergraduate level and at post-qualifying and Masters’ level, in keeping with requirements of professional bodies. These courses are offered both on a full time basis and part time for qualified practitioners working in the field accessing continuous professional development (CPD). Evaluation of these courses suggests that some students find research methods challenging to understand and the pace of sessions demanding, and has highlighted a need for additional ways to support learning and teaching. There are a number of existing electronic resources relating to research methods accessible to students via the internet, which could help to support learning and teaching in this area and meet the wide range of learning styles among students. However, many are not specific to health research. In addition, the quality of content can be variable and use/accessibility unpredictable. This, combined with the need for innovative ways to engage interest in research methods, suggested the need for a new electronic resource for health research, for use within the context of a classroom taught course. The process of developing an interactive resource incorporating a narrative element is described. A narrative approach recognises the power of story in capturing interest and transferring information and offers scope for imagination and intrigue within learning. A story of two fictional health practitioner characters working in a local health centre was created to weave around research methods theory. Interactive elements such as question-and-answer tasks, audio extracts, games and interactive graphics were added to offer varied and stimulating ways of presenting material to meet a range of learning styles. The resource also incorporates a number of self-assessment opportunities to reinforce learning. The use of voices heard in realistic scenarios arising in the health centre anchors learning in everyday practice aiming to help students appreciate the need for evidence and the value of research understanding.

Keywords: research methods teaching, evidence-based practice, elearning, nurse education, narrative

1. Introduction

The Using Health Research project was funded by the Department of Nursing and Midwifery at the University of the West of England, Bristol (UWE), a major provider of health and social care education in the south west of England and one of the largest in the UK. The aim of the project was to meet a need for an electronically based, interactive learning resource in research, a core component of the nursing curriculum (NMC 2004), for use within a module delivered to post-registration health professionals. It was carried out between October 2008 and February 2010 in collaboration with an external partner, whose previous work in developing a resource for teaching politics research had inspired the initial idea (Middleton & Bridge 2008).

This paper presents an account of the development of the web-based and DVD resource, designed for use within a specific research module at UWE. This paper offers an insight into the process of designing and developing the resource, and a perspective on the potential for this medium in research teaching more widely in the Department. It is also an exploration of the use of story in teaching research and in this sense represents a novel approach to this area of the curriculum which is commonly taught in lectures with seminars based on critiquing published research articles. At the time of writing, the evaluation of the resource is ongoing hence this element of the project cannot yet be reported.

2. Background

2.1 Nurse education

A major challenge in nurse education is the high volume of students. In 2008/09 registration on Higher Education (HE) nursing courses outranked all other HE courses by more than 40,000 (HESA 2010). This presents just one of several drivers for innovation in pedagogy in the field. Wider participation in higher education has brought larger and more diverse student groups and in turn, challenges ranging...
from practicalities such as space and seating, to noise and comfort but also to pedagogical demands in delivering content in ways that meet the inevitably wide variety of learning needs and styles present in the classroom.

Alongside this, the current generation of students bring expectations of greater service and convenience (Garrison and Kanuka 2004) and of increasing use of technology in teaching and learning (Brown et al 2008), offering greater mobility, self-direction and interactivity. One response to these challenges is an expansion of elearning approaches in the field of nurse education which McVeigh (2009) asserts is essential to future educational development and the facilitation of life long learning, in keeping with wider policy objectives. Although there is rightly some caution around the promotion of electronic modes of learning (Atack, 2003, Farrell 2006, McVeigh, 2009) there is considerable appeal to those responsible for workforce development who are facing continual staff shortages and reduced training budgets making it harder to release staff from clinical environments (Moule et al 2008), which may further drive demand in this area. More widely technology-based practices are perceived to have the potential to transform HE in the 21st century. (Garrison and Kanuka 2004)

2.2 Research and evidence-based practice

A compulsory element of the nursing curriculum is the teaching of research and the role of evidence in practice, starting in pre-registration education (Department of Health 2006), and required by the Nursing and Midwifery Council which governs nurse education in the UK (NMC 2004). Given the high numbers of nursing students referred to above this results in a large number of modules in health research across pre-registration, post-registration and Master’s level courses and thus a need to find imaginative ways to teach a complex subject at a variety of levels.

The module for which the resource was developed is a post-registration module currently taught using face-to-face sessions with a mix of didactic teaching, individual and group activities intended to help apply the ideas. As a core component of the curriculum research is a compulsory module ensuring full classrooms throughout the year. Students are traditionally in full-time employment in a health or social care role, usually nursing and midwifery but more recently the course has seen popularity with emergency-care professionals, bringing a mix of health service workers together for the course. Students attend one day per week for eight weeks, with two half-days free study time out of this timetable. Library and IT support are provided within this programme and access to course tutors. Students have little spare time to prepare for sessions or for individual study during the eight weeks and find the course intensive and challenging due to the technical nature of the subject, which few have studied before. They often express a need for ways to reinforce learning and at times convenient to working people. Although there are online resources to support research teaching/learning available on the web, the authors were not satisfied with those available at the time of the project, either in terms of accessibility or relevance to health and social care research or a match with course objectives. A key objective in developing a new learning resource was therefore to support learning outside the teaching sessions, with flexible access, in a way that complemented the materials used on the module and the students’ learning context.

The greatest challenge, however, was perhaps the need to make this resource engaging, interactive and fun. Many students in nursing express the view that research is a difficult subject (Owens and Kelly 1998) and experience of teaching research bears this out, as well as revealing the struggle among learners to find research interesting, bringing considerable pressure for educators to develop stimulating and creative ways to teach topics such as sampling and statistics. Previous research in the field of research teaching has shown that an internet-based approach can result in better learning stimulation (Woo and Kimmick 2000) and more recently that a blended approach can be beneficial to the nurse learner (Johnson et al 2009). Other areas of the nursing curriculum have also found success with multimedia approaches and have shown how this can help meet individual learning needs and lead to a greater sense of control over learning and of self-achievement compared to traditional lectures. (Moule 2002)

Classroom experience further highlights the importance of engaging students during teaching sessions with the application of research to professional practice, leading to appreciation of the relevance of research to personal experience and thus to deeper learning (Kolb 1984). This may seem obvious and is, ultimately the point of teaching practitioners about research in the first place, but can be easily lost when the course focus is on understanding research rather than doing research,
typical of courses at this level. An approach used by one of the authors to achieve this has been to position a fictional patient in the learning materials such that research examples that are appraised are discussed in relation to this patient and their care needs. This has proved popular with students and further fuelled the idea of using story in research teaching.

2.3 Narrative and the use of story in learning

Within nursing education, there is an emerging field of narrative pedagogy which as Brown et al (2008) explain, is an approach that incorporates the use of art, film, music and literature in teaching which can bring more powerful and sensitive insights into illness and caring. Alongside traditional storytelling this approach has seen the use of personal narratives and stories to support reflective learning which is a core element in nursing education. McDrury and Alterio (2003) explain that storytelling provides students with the opportunity to engage in reflective dialogue through which reflexive capacity can develop and is gaining popularity in higher education more generally, perhaps because it incorporates elements of reflective, experiential and constructivist models of learning that inform adult education. It is not the intention here to argue the case for narrative pedagogy but to acknowledge a debt to this approach: firstly and most simply, the celebration of storytelling and recognition of the power of story in transferring information, to “capture interest and attention…and bring facts to life by putting them in personal scenarios” (Brown et al 2008:284), echoing Bruner's (1996) notion of ‘narrative thinking’. Secondly, in seeking to connect learners and educators through shared stories and thirdly, in promoting reflection in learning. (Brown et al, 2008; McDrury and Alterio, 2003)

These were the foundations of the idea to create a ‘soap opera’ style narrative to weave through the teaching resource with the intrigue of a story intended to carry the learner through the theory elements. The original conception was to see characters in a story set within a primary health service setting, giving opportunity for representation of multidisciplinary roles as well as reflecting the current shift in UK health service policy to expansion of community-based services, and this remained unchanged. In addition, the particular context of the soap story within a health service setting was intended to capitalise on the great popularity of medical and hospital dramas on UK television.

3. Phase 1: The project team

The first phase of the project was the formation of a project group and period of consolidation. The team had not worked together before but the authors had established a common interest with the external collaborator at a conference, leading to discussions about developing a resource for health research. Agreement was reached to work together, with the collaborator bringing a consultant from the previous project to support software development, and a joint project was set up in 2008. The team brought together three distinct disciplines in addition to teaching and learning design: primary care nursing, health and social research, and politics research. There was a strong commitment by the four members of the group to the goal of producing a high quality teaching resource but the different disciplines represented in the group made it essential to spend the early part of the project reaching consensus on how to work together and manage the project but also more importantly, consensus on a vision for the project.

A period of ‘storming’ and ‘norming' typical of a new group (Tuckman 1965) followed. Balance was negotiated between the strengths that each group member brought to the project – on the one hand knowledge of professional practice from the health and social care context – and on the other teaching politics research and production of elearning resources, in order to agree project tasks and leadership. Although this mix generated creative synergy, an area of complexity was navigating the roles of commissioner and client with responsibility for overall good financial governance and time management whilst also taking a full part in design and authorship.

A potential dissonance related to precedence of professional perspectives. In the interests of the narrative more dramatic events or story were suggested but the need to reflect professional standards within nursing and other professions demanded greater realism within the stories and characterisation. Furthermore the group were mindful that the end product would be associated with an academic institution. An additional challenge for the group arose from the geographical distance between all four team members which led to managing the project mainly through email with only monthly or less frequent face-to-face meetings. The remote nature of communication certainly slowed
the process of team building and trust, and compounded any potential miscommunication, in turn creating delays in the project timetable.

4. Phase 2: Design of the resource

4.1 Learning content

Early meetings focused on clarifying an overview of all the tutorials in terms of the learning aims and objectives. Ideas were shared about options for how to present material and the team drew on their experiences of teaching in the classroom and discussed what would and would not transfer effectively to an electronic mode of learning. The need to meet a variety of learning styles among adult students of varying backgrounds, and needs relating to accessibility, were also discussed. The initial concept was to provide six tutorials covering research design and methods with four of these to be completed in the first year. At this stage it was anticipated that some of the theoretical content and activities from the collaborator’s previous resource could be adapted to form material for tutorials in Using Health Research. However, this later proved impractical as the design of tutorials developed and the product evolved. This change, and the decision to develop all six tutorials in parallel for reasons of economy caused considerable pressure for the team.

The decision was taken to design the resource as a linear sequence of tutorials leading the learner on a pathway through the materials and unfolding story, although it could be used more loosely as a reference resource. The story itself was an archetypal ‘boy meets girl’ tale of workplace romantic intrigue in the context of a busy health centre, with the sub plots dominated by the pressing demands of frontline service delivery as delegated by the practice manager and a senior GP. It was originally envisaged that a key character in the story could be chosen by the learner and followed through the various topics giving an opportunity to identify with an individual and their role. This was later altered to having two main characters in parallel in the storyline with learners seeing both involved in different scenarios to make a clearer route through the tutorials. Alongside the discussion of learning content, the idea of professional ‘dilemmas’ emerged. A dilemma would be a practical problem encountered by characters in the story to help the learner adopt a problem-solving stance within the resource bringing further motivation to follow the theory elements to solve the dilemma. This approach would also serve to model the use of evidence in practice.

The development of the first tutorial took longer than planned. Considerable time at the first three meetings was given to reaching a shared understanding of the module and the specific elements relating to health research and evidence-based practice, to agree content and how the resource would fit with the face-to-face teaching as in a ‘blended learning’ approach (Bonk and Graham 2006). This determined the need for revisions to subsequent tutorials, leading to changes in content relating to research design and methods. The importance of not repeating content from the face-to-face teaching sessions in detail, was a key factor at this stage; material presented should refresh knowledge and develop understanding through opportunities for interaction with material and self-assessment to support students’ learning. This put pressure on finding innovative ways of presenting material and an emphasis on interactivity where text would be in small chunks such as in a quiz or other question-and-answer formats interleaved with the presentation of the story element. Later on in the process, the design of the second tutorial showed that too much content was being presented such that it became necessary to have a separate tutorial on sampling. This necessitated the decision to postpone a tutorial on ethics for a second project.

An early online prototype was produced to illustrate the outline structure and options for navigating through the resource, which helped the authors with design of content for presentation on screen and the team as a whole with deciding layout and style, section headings, hyperlinks and menus. The final product includes a welcome page with introduction to the resource and the characters within the story. Each screen has a bold black header showing an ECG trace of a healthy heart rate alongside the title, and below this a row of brightly-coloured tabs for each tutorial, numbered and named. The six tutorials are: 1. Ways of knowing (covering concepts relating to types of knowledge and evidence). 2. Design and methods (including an overview of epistemology and the assumptions within qualitative and quantitative approaches, with activities relating to popular designs and methods in health research). 3. Sampling (with comparison and illustration of probability and non-probability methods). 4. Evaluating research I (questioning issues of validity and reliability in different designs and methods). 5. Evaluating research II (continuing from tutorial 4). 6. Working with data (which introduces ways to begin the interpretation of research data from different approaches). Each tutorial has up to
seven sub-sections, with buttons marked ‘next’ to direct the learner through the tutorial with the option to use a ‘back’ button to return to the previous page, with the coloured tabs for jumping directly between tutorials. At the top of each screen hyperlinked text advises which tutorial and sub-section the learner is currently viewing offering another way to navigate to previous sections.

4.2 Characterisation and writing the soap opera

Initially, ideas about the storyline were discussed in parallel with the learning design and content to help clarify the product as a whole. At the second meeting, it was decided that an introduction to the characters should be written to initiate characterisation and story-writing. Two fictional characters were devised as the two main people in the story. After much deliberation about roles, age, gender, experience and personality they emerged as a male health visitor, in his late 20s, newly qualified and keen, with an interest in research. The second was a female nurse practitioner, in her early 30s, more experienced but new to the health centre. The characters were named and a narrative about their lives and the wider primary care team and health centre evolved through discussion in the team, including a vision of the town in which the health centre was located and its demography. The names used in a first draft were kept except for the main female whose name was later changed to be more fitting with the character that had evolved. With names, the characters began to take on greater realism and group discussion at the face-to-face meetings was an important part of developing personalities for the characters and features of the story.

The first scripts began to be written in the second month of the project and the dilemmas were drafted. Script writing required blending of different styles and ideologies of the team members. The importance of realistic use of language and credible interactions between colleagues was the subject of much debate to balance the dramatisation with current primary care policy and practice. A subtlety of this was navigating the strands of entertainment and learning so that the story did not distract from the theoretical content. Script-writing extended over four months. Highly constructive editing ensured scripts and storyline were refined benefiting the credibility of the final storyline.

4.3 Audio-visual elements

The budget determined that the soap opera would have to be presented using audio recordings as the cost of video production was prohibitive. Originally it was envisaged that this aspect of the production might be created within the university, however the timescale precluded this option and the production team previously used by the collaborator was utilised. Professional actors played the characters in the scripts and a small production team comprising a sound recordist and media producer managed the highly professional and cost effective production. The six actors met for a day to read-through scripts together before the final recording. Hearing the written words come alive with actors’ voices was hugely encouraging for the group at this stage. It also gave the opportunity to refine phrasing and advise on pronunciation of technical terms which enhanced the professionalism and quality of the final production.

Ideas for incorporating external video such as extracts from television programmes to bring further variety in the presentation of material had to be abandoned due to copyright costs.

4.4 Interactivity and learning activities

The commencement of script-writing and dilemmas for the characters initiated discussion of activities. This capitalised on the expertise in the team in designing e-learning activities. The audio offered students a means to engage with and reflect on the ideas raised, being a key element of learning through storytelling (McDrury and Alterio 2003). However, additional activities and tasks were felt to be important in giving learners the opportunity to think more deeply about the concepts and importantly, to use them in ways that tested and developed understanding. Regular self-assessment activities should also help learners identify the strengths and weaknesses in their knowledge and understanding. Initially the team expected to incorporate various links to external resources such as research articles and other websites for example to view data or research tools such as questionnaires, but as the design evolved these options were abandoned, partly to avoid distracting the learner away from the resource but also to have greater control over the material presented. The range of tools in the final version of the resource includes: considering scenarios and ticking responses from a list; drag-and-drop activities to test knowledge and understanding such as definitions, simple concepts and material from the audios; a quiz presented in such a way that each
correct answer exposes a section of picture that the learner may seek to complete; entering text into forms and checking responses; interactive graphics where areas are selected and explanatory text pops-up; animated graphs to illustrate data and consequences of different options; simple answer selection with feedback; a fruit machine with winning payout for correct selection of terms. In addition to the necessary text around the activities throughout all the tutorials the learner is directed to listen to the short sections of audio with many of the activities linking to content heard. This variety in interactivity helps to meet preferences for learning through listening, reading, reflecting, decision-making and application.

4.5 Image

The importance of visual images emerged early on in the project with sharing of pictures and ideas about how the resource should look. The look and feel of the resource needed to reflect the setting of a local primary care health centre with images and characters to capture the atmosphere and business of such a setting. Ambitions were limited by resources as the group would have favoured a dynamic, audio-visual introduction, rather like the credits to a TV programme showing a wider setting of the health centre (town scenes with people, roads, homes, shops) accompanied by music, and then moving to close-up scenes of people arriving and entering the centre. In the initial stages the importance of not getting distracted by the design side of production was recognised and these ideas stayed at the discussion stage until more content was written.

A later phase of choosing visual images related to the presentation of the main characters in the soap opera. Images were needed to accompany audios to help the listener recall the speakers and contextualise the conversation. Also images would form part of the welcome page to introduce the characters, setting and location. As the copyright for so many photographs was too expensive, an artist was used to produce drawings that could be situated within photographic backgrounds to give the appropriate context and the variety of images needed. Background scenes were photographed by the authors at locally based health centres with their permission.

5. Phase 3: Implementation

Implementation of Using Health Research has recently commenced within the module for which it was designed. This involves providing students with the web address for accessing the resource and in the near future, they will be offered a DVD version that can be accessed offline to increase use among those without readily available internet access. Students can use the resource independently but will be directed to appropriate sections or activities within the website at the end of lectures to remind them of the opportunity for further interaction with the ideas and self-assessment. It is anticipated that lecturers may also wish to use some of the quiz-style activities within sessions. An evaluation among the students on the module is being planned. The resource was demonstrated at a Departmental advisory group for blended learning and received much positive comment. The Department are now intending to evaluate the resource for use within other research modules within pre-registration education.

6. Limitations

As with many projects it is possible to see the caveats before the first version is complete. The use of video instead of audio would have added greater engagement with the soap opera element and added more visual stimulation whilst listening to the conversations. However, it is conceivable that students might have been distracted by too much visual information relating to the story and not therefore fully appreciated the learning points about research. Although the modelling of the scenarios in relation to learning about research methods was accurate in relation to methodology, the team were aware that the complexity around undertaking research in a health care setting was not fully portrayed, especially around the requirements for ethics and research governance. A tutorial on ethics was cut to avoid overloading other tutorials with theory though this was partly rationalised by the assumption that a second project could follow to incorporate other topics not covered in the first version. Inevitably the scripts were, to an extent, driven by the need to give context and deliver research learning material at the expense of developing characterisation or story leading to some stereotyping at times. Enthusiasm for the story, an eye for dramatic detail and pace at times led the authors into internal debate where these posed questions of professional propriety. As a result certain details of the scripts are the product of three opinions which did not always converge, making some lines and story features disjointed. A further issue in relation to the production of a resource that attempts to contextualise learning by using a familiar setting and characters is the pace of change in
health and social care delivery and services management itself, threatening the currency of the context portrayed here, and necessitating regular review and updating. However, like any good story the possibility of a sequel has been built in and a further edition would be a possibility.

7. Conclusion

The development of Using Health Research has been a significant undertaking requiring huge commitment by the authors and taking over 18 months to complete. The work was made more challenging by the geographical distance between team members and concurrent workload of all involved and it is possible that without these hurdles the product would have been further enhanced. Nonetheless, the commitment has produced a novel, creative, high quality resource that helps to meet the range of learning needs and styles of the diverse student population, and with fairly modest funding. It is hoped the evaluation will reveal this to have been well spent in terms of the effectiveness and popularity of the resource, and potentially the value as a pilot product with wider marketability. The decision to roll-out the resource to other research modules will need to be underpinned by robust evaluation testing the effectiveness of the resource with other learners, but the product may have value beyond the initial intention.

Acknowledgements

The authors gratefully acknowledge the support for project funding from the Faculty of Health and Life Sciences, University of the West of England and the expertise of our collaborators Dave Middleton and Geoff Austin. We would also like to thank the Gloucester Road Medical Centre and the Old School Surgery, Bristol for giving permission for photographs of their premises to be used.

References

A Different Vision in eLearning: Metaphors

Nazime Tuncay¹, Ioana Andreea Stanescu², and Mustafa Tuncay³
¹Near East University, Nicosia, North Cyprus
²Advanced Technology Systems, Romania
³Atatürk Teacher Academy, North Cyprus
nazime.tuncay@gmail.com
ioana.stanescu@ats.com.ro
mustafa.tuncay@aoa.edu.tr

Abstract: Metaphors are figures of speech in which a word or phrase that denotes a certain object or idea is applied to another word or phrase to imply some similarity between them. Due to their ability to make speaking and writing more lively and interesting, metaphors have always been popular among students. While metaphors provide significant enhancement of contexts and build upon the sense of community, they can limit the boundaries of the communication between students and teachers. In order to carry out student oriented courses, teachers ought to consider the metaphors students use. In an effort to understand and fill in this communication gap, the authors of this paper have initiated a study that aimed to drive out the e-education students’ metaphors in order to suggest a vision for future e-courses. The authors have designed the “E-Education Metaphor Analysis Survey” that comprised 35 items and captured data about e-education students’ metaphors. The questionnaire was posted on Surveymonkey.com and was distributed to e-education students in two countries: Turkey and Cyprus. 352 students filled the questionnaire. The answers revealed that the metaphors students use are influenced by their way of life, their personal characteristics, their educational background and their feelings. Internet was the most common metaphor used for e-education. A very interesting fact was that 47% of the students considered E-Student to be equivalent to “rich students’ education” and that the term recalled them the metaphor “richness”. Although there were many research studies on common metaphors and their impact on e-education, there were no studies in the literature about eLearning metaphors. This paper presents an innovative approach that focuses on 7 key research questions and represents a first step of a more detailed future project undertaken by the authors.

Keywords: eLearning; metaphors; students; SurveyMonkey

1. Introduction

Students learn in different ways, they develop their own understanding patterns, and they build new communities and new language items that grow more and more complex and that can prove really challenging for teachers. To reach a common ground, teachers need to constantly consider new perspectives that could lead to the enhancement of the communication with their students, and consequently of the learning process. A very practical approach is to take note of the metaphors that students use in their daily life.

Metaphors sustain a better understanding of an object or of an idea to which the metaphor is being applied and they can communicate a more accurate meaning with just a word or a phrase. More, they create a sense of belonging and acceptance in certain communities, strengthening group communication. Learners are most satisfied with courses in which the instructors facilitate frequent contact between themselves and learners, where they use active learning techniques, convey high expectations, emphasize the time spent on specific tasks, and provide prompt feedback (Polloff and Pratt, 2001). Under these premises, metaphors present a great potential for the highly complex learning process and their usage can impact positively upon the students’ results.

1.1 The term eLearning and eLearners

Although numerous definitions of eLearning are found in the professional literature, this term often refers to instruction or training delivered using media, computers, and technologies such as the WWW and Intranets. The content delivered via eLearning is related to (a) instructional goals, (b) specific instructional methods, (c) selected media, and (d) knowledge and skills for achieving individual or organizational goals (Uzunboylu, 2007; Muthukumar, 2004; Clark and Mayer, 2003).

There is a growing body of literature on eLearning technologies (see for instance Gayeski, 1993; Gibbons & Fairweather,1998;Kearsley, 2005; Khan, 1997);

Learners have different ways of learning and their issues such as instructional support, faculty motivation and enthusiasm, and technology problems have been raised as problems in developing
online instruction in many institutions for a long time (Barr and Tag, 1995). There may be several issues effecting learners’ perceptions about online learning: Instructor, Website, Computer Skills, Pedagogical Issues and English Language. If learners are not satisfied with the design of the course website, they may have negative perceptions of the effectiveness their online courses (Brush, 2001). Technological problems and pedagogical issues make learners get frustrated with the online courses (Hara and Kling, 2003).

1.2 The impact of metaphors in eLearning

Metaphors have proven to be a highly useful tool in the development of theories in the social sciences (Hartzell, 2004; Kendall & Kendall, 1993; Levassuer, 2004; Wang, 2004). They provide a convenient means by which to create a taxonomy; the first step towards description, then prediction and finally to understanding (Kerssens-van-Drongelen, 2001; Lewis &Grimes, 1999; Lynham, 2000). Also, metaphors become essential elements that comprise the everyday language among specialists (Cook-Sather, 2003) and they are used to communicate meanings without describing them directly but using some element that could help others understand the concept individuals want to communicate.

There are many metaphors that surround the concept of eLearning. Each of them provides some insight and represents a different understanding, but taken together they reveal what a complex concept eLearning really is. Each metaphor offers unique perspectives for the group that is using it and, at the same time, limits understanding in various ways both for the group that has generated and is using the metaphor and also for people residing outside of the group. This fact constitutes premises for broadening the gap between teachers and students and disagreements may be just a result of each group talking past each other while using different metaphors.

Metaphors embody and define the intangible and abstract, but this process inevitably constrains perceptions and actions to those which make sense within the logic of the metaphor. Metaphors are therefore both descriptive and prescriptive. It is important that both students and teachers become aware of their own metaphors for eLearning, so they can recognize how these limit or liberate them. Working together with other learners increases involvement in learning and deepens understanding (Chickering & Ehrmann, 1996). These differences in students’ perceptions and understanding result in them developing and using different metaphors, for every metaphor highlights one aspect of the concept, just as it hides another; Lakoff and Johnson (1980) call this “metaphorical systematicity”. In his analysis of McLuhan’s impact, Levinson (2001) reflects his regrets that McLuhan’s statements have fueled “the fire of worry that bad things are happening that we can’t know or understand”. And while McLuhan proposed that the “medium has an impact above and beyond what we do with it” (Levinson 2001), there is no firm evidence as yet that the worriers are correct (Meyer 2002). Tuncay&Özçınar (2009) have carried out a research study on 106 university students and reached the conclusion that for E-class the Paris metaphor was chosen by 22.2%; for Distant Student, "computer heroes" metaphor was the most commonly chosen; for Blended education, the metaphor "Girls and Boys together" was the most commonly chosen (22.2%); for Synchronous Education, "Morning Education" was most commonly chosen (33.3%); for Asynchronous education, computer metaphor was the most commonly chosen. Students’ metaphors have great impact in their life (Tuncay&Uzunboylu, 2009) and teachers must be aware of the metaphors their students use, in order to carry out their courses as student oriented.

1.3 Purpose of the study

Most of the research studies in the literature were related to common metaphors and their impact on distance education (Meyer, 2005) and none of them approached e-education metaphors. The main purpose of this study is to drive out the university students’ metaphors, build upon a better communication among students and teachers, and suggest a vision for future e-courses.

2. Method used

2.1 Population

Over 600 questionnaires were distributed exclusively online via email, in Turkey and in Cyprus. When the submission was closed, after 1 month, a total of 352 questionnaires were collected. The
distribution of the questionnaire to the target group has proved a real challenge, and generally each student was questioned two times.

2.2 Instrument

To capture data about university students training needs, the authors have created the “E-Education Metaphor Analysis Survey” questionnaire, consisting of 35 items, and they have distributed it online. The online questionnaire was designed (see Figure 1) and posted (see Figure 2) on SurveyMonkey.com and it has been emailed to university students that have previously been taken e-courses. In the first phase, the questionnaire has been distributed in two countries (Turkey and Cyprus), in Turkish. In order to evaluate the items in the questionnaire, expert evaluation (n = 13) was required. Experts group from education technologist evaluated the data gathering scale both individually and collaboratively, and sustained the maintenance of the content’s validity. All the experts’ evaluations and suggestions were taken over in the draft form of the questionnaire and afterwards the necessary corrections were made; some of the questionnaire questions included: what comes to your mind when one mentions E-education? What comes to your mind when one mentions E-course? What comes to your mind when one mentions E-counselor? What are 10 most favorite metaphors that you come across in your e-courses?

2.3 Data analysis

The responses were downloaded from the surveymonkey.com. The authors considered descriptive statistics as a means to present quantitative descriptions in a manageable form and reduce lots of data into a simpler, yet more powerful summary that enables comparisons across people or other units. Descriptive statistics frequencies and percentages were used to analyze and to report the data gained from the questionnaire SurveyMonkey, as frequency and percentage statistics are considered more relevant in the representation of personal information variables. Frequency represents the number of participants who indicated that category (aka "Male"), while percentage was calculated by taking the frequency in the category divided by the total number of participants and multiplying by 100%. The study reports descriptive statistics such as gender, age and country.

![Designing questionnaire on SurveyMonkey](image)

**Figure 1:** Designing questionnaire on SurveyMonkey
3. Results and discussion

In this section, the authors present in detail the survey results by item, respectively: eLearning metaphors, country, gender, percentages of e-metaphors, and common metaphors, identified as relevant in this effort to fill in the gap and build the grounds for common understanding between teachers and students.

3.1 eLearning metaphors

The metaphors derived by students and their percentages are (see Figure 2):

- TV Channel (0.3%),
- Human Brain (0.6%),
- Java Games (1.7%),
- Other (4.9%),
- Magazine (11.5%),
- Classroom (17.5%),
- Digital Story (18.9%) and
- Internet (44.7%)

The spectrum of options reflects the evolution of education from traditional to technology-based approaches, allowing respondents to situate themselves in a non-ambiguous perception zone.
3.2 Responses according to country

The survey was distributed in Turkey and in Cyprus. Since the two cultures share common backgrounds, no significant differences were seen in the answers to the question “What comes to your mind when one mentions about asynchronous education?” (See Figure 3).
3.3 Responses according to gender

60 of the students have not specified their gender. Overall, 136 men and 140 women have participated in this study. An independent sample t-test found no significant difference between groups (p>0.05) which have specified their gender.

For asynchronous education, the students from Turkey have chosen the “morning classes” metaphor in a greater percentage than the students from Cyprus, while the students from Cyprus have selected the “Classroom”, “Light” and “Non-technological” metaphors at a greater percentage than the students in Turkey.

3.4 Percentages of “E-” Metaphors

Table 1 lists the e-education metaphors, which have the highest ranks, based on the questionnaire results. According to these, the E-school students said that “I feel like I am wondering in space”. The E-student should go after information as an e-detective. The E-teacher is the one who has the last word. 47% of the students answered that the E-Student is equivalent to “rich students’ education” and that the term recalled them the metaphor “richness”. Their reasoning was: “You need to be rich to have education via mobile phone”. 48.3% of the students paired E-Project to “Imaginary World” and their reasoning was that objects generally are never written on paper, it is like in an imaginary world”.

Table 1: E-metaphors

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Metaphors</th>
<th>Reasoning</th>
<th>Percentage of the Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-Class</td>
<td>E-Group</td>
<td>There are lots of e-groups that use Skype, Wikipedia and so on to share information with classmates</td>
<td>85.7%</td>
</tr>
<tr>
<td>E-Teacher</td>
<td>Judge</td>
<td>They judge our work and take decisions</td>
<td>71.4%</td>
</tr>
<tr>
<td>E-Quiz</td>
<td>Hot-potatoes</td>
<td>We come across lots of quizzes created by hot-potatoes</td>
<td>68.9%</td>
</tr>
<tr>
<td>E-Counselor</td>
<td>Distant Relative</td>
<td>They are like distant relatives they say what to do without exactly being in the same conditions, even in the same country, more like distant relatives</td>
<td>60.0%</td>
</tr>
<tr>
<td>E-School</td>
<td>Space</td>
<td>We can not physically be in it</td>
<td>58.3%</td>
</tr>
<tr>
<td>E-Help</td>
<td>MSN</td>
<td>We generally get e-help through MSN</td>
<td>56.0%</td>
</tr>
<tr>
<td>E-Student</td>
<td>Detective</td>
<td>They should do lots of researches and search for clues of success like detectives</td>
<td>51.4%</td>
</tr>
<tr>
<td>E-Project</td>
<td>Imaginary World</td>
<td>Projects generally are never written on paper, it is like in an imaginary world</td>
<td>48.3%</td>
</tr>
</tbody>
</table>

3.5 Common metaphors

“What are 10 most favorite metaphors that you come across in your e-courses?” was one of the questions in the questionnaire. The students’ answers and reasoning were consistently different (listed in Table 2). The words which have highest metaphor rank (85.7%) are: Athlete (metaphor: Greyhound); Guitar (metaphor: Antonio Banderas); Mother (metaphor: Water); Sea (metaphor: Bed cover); University (metaphor: Money). The other words which have highest metaphor rank are: Cloud (metaphor: Cotton); Red (metaphor: Blood); Clock (metaphor: Water); Poplar Tree (metaphor:Tall); Comb(Hairdresser); Ring(metaphor: Marriage). There are other metaphors that have been used, reflecting the diversity of approached that students have been taken and also they vivid imagination that triggers a constantly open challenge for their teachers.
Table 2: Commonly used metaphors

<table>
<thead>
<tr>
<th>Noun</th>
<th>Metaphor</th>
<th>Common Reasoning</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athlete</td>
<td>Greyhound</td>
<td>Both are agile and quick</td>
<td>85.7%</td>
</tr>
<tr>
<td>Guitar</td>
<td>Antonio Banderas</td>
<td>Antonio Banderas plays guitar in films</td>
<td>85.7%</td>
</tr>
<tr>
<td>Mother</td>
<td>Water</td>
<td>She is pure and beautiful as water and we need her to be with us as we need water</td>
<td>85.7%</td>
</tr>
<tr>
<td>Sea</td>
<td>Bed cover</td>
<td>They both makes you rest</td>
<td>85.7%</td>
</tr>
<tr>
<td>University</td>
<td>Money</td>
<td>You have to have money to go to university</td>
<td>85.7%</td>
</tr>
<tr>
<td>Cloud</td>
<td>Cotton</td>
<td>Same visuals</td>
<td>84.6%</td>
</tr>
<tr>
<td>Red</td>
<td>Blood</td>
<td>It’s the first thing which comes to our mind.</td>
<td>84.6%</td>
</tr>
<tr>
<td>Clock</td>
<td>Water</td>
<td>Clock goes fast and it does not come back. Similarly when water falls down it does not come back into its container.</td>
<td>76.9%</td>
</tr>
<tr>
<td>Poplar Tree</td>
<td>Tall</td>
<td>Poplar Tree is tall</td>
<td>76.9%</td>
</tr>
<tr>
<td>Comb</td>
<td>Hairdresser</td>
<td>Hairdressers use comb too often</td>
<td>75.7%</td>
</tr>
<tr>
<td>Ring</td>
<td>Marriage</td>
<td>Ring is a symbol of marriage</td>
<td>75.7%</td>
</tr>
<tr>
<td>America</td>
<td>Freedom</td>
<td>America is a place of freedom</td>
<td>71.4%</td>
</tr>
<tr>
<td>Gamble</td>
<td>Cyprus</td>
<td>People come to Cyprus to gamble</td>
<td>71.4%</td>
</tr>
<tr>
<td>Noun</td>
<td>Metaphor</td>
<td>Common Reasoning</td>
<td>Frequency</td>
</tr>
<tr>
<td>Leg</td>
<td>Column</td>
<td>Both are straight and long</td>
<td>71.4%</td>
</tr>
<tr>
<td>Scissors</td>
<td>Tailor</td>
<td>Tailors use scissors too often</td>
<td>71.4%</td>
</tr>
<tr>
<td>Turkish people</td>
<td>Wolf</td>
<td>Wolf is a symbol for Turkish people</td>
<td>71.4%</td>
</tr>
<tr>
<td>Eyelash</td>
<td>Türkan Şoray</td>
<td>Türkan Şoray has long and beautiful eyelashes</td>
<td>68.9%</td>
</tr>
<tr>
<td>Ice</td>
<td>Glass</td>
<td>You can see through them both</td>
<td>68.9%</td>
</tr>
<tr>
<td>Summer</td>
<td>Sea</td>
<td>We swim in the summers</td>
<td>68.9%</td>
</tr>
<tr>
<td>Distance Education</td>
<td>Mars</td>
<td>It makes us believe that we can have courses even in the Mars.</td>
<td>60.0%</td>
</tr>
<tr>
<td>Horse</td>
<td>England</td>
<td>In films English people ride horses</td>
<td>60.0%</td>
</tr>
<tr>
<td>Moustache</td>
<td>Kadir İnanır</td>
<td>Kadir İnanır is a famous artist with moustache</td>
<td>60.0%</td>
</tr>
<tr>
<td>Spain</td>
<td>Bull</td>
<td>Spain is famous with bull fights</td>
<td>60.0%</td>
</tr>
<tr>
<td>Woman</td>
<td>Flower</td>
<td>They love flower</td>
<td>60.0%</td>
</tr>
<tr>
<td>Holiday</td>
<td>Antalya</td>
<td>It’s a favorite Turkish holiday place</td>
<td>58.3%</td>
</tr>
<tr>
<td>Mushroom</td>
<td>Umbrella</td>
<td>Very similar</td>
<td>58.3%</td>
</tr>
<tr>
<td>Snow</td>
<td>Erzurum</td>
<td>Erzurum has too much snow</td>
<td>58.3%</td>
</tr>
<tr>
<td>Walnuts</td>
<td>Brain</td>
<td>They resemble in shape</td>
<td>58.3%</td>
</tr>
<tr>
<td>Chicken</td>
<td>Döner</td>
<td>Chicken meal</td>
<td>56.0%</td>
</tr>
<tr>
<td>Dog</td>
<td>Smell</td>
<td>Dogs are sensitive to smells</td>
<td>56.0%</td>
</tr>
<tr>
<td>Human Being</td>
<td>Monkey</td>
<td>Dawin’s theory, we, human beings developed from Monkey.</td>
<td>56.0%</td>
</tr>
<tr>
<td>Needle</td>
<td>Hay</td>
<td>They resemble</td>
<td>56.0%</td>
</tr>
<tr>
<td>Pizza</td>
<td>Italy</td>
<td>Italy is famous with its pizza</td>
<td>56.0%</td>
</tr>
<tr>
<td>Student</td>
<td>Slave</td>
<td>Must do whatever exercises the teacher gives</td>
<td>56.0%</td>
</tr>
<tr>
<td>Blue</td>
<td>Sky</td>
<td>The blue is the color of the sky</td>
<td>52.9%</td>
</tr>
<tr>
<td>Phone</td>
<td>Mobile Phone</td>
<td>When somebody says phone we understand mobile phone</td>
<td>52.9%</td>
</tr>
<tr>
<td>Airplane</td>
<td>Bird</td>
<td>Both are in the sky</td>
<td>51.4%</td>
</tr>
<tr>
<td>Fork</td>
<td>Rake</td>
<td>They resemble in shape</td>
<td>51.4%</td>
</tr>
<tr>
<td>Lahmacun</td>
<td>Gaziantep</td>
<td>Lahmacun is famous food in Gaziantep</td>
<td>51.4%</td>
</tr>
<tr>
<td>School</td>
<td>Home</td>
<td>School is warm as a home</td>
<td>51.4%</td>
</tr>
<tr>
<td>Tea</td>
<td>Rize</td>
<td>Rize has a different taste of tea</td>
<td>51.4%</td>
</tr>
<tr>
<td>Father</td>
<td>Mountain</td>
<td>Is powerful. No danger comes to you as far as your father is near you.</td>
<td>48.3%</td>
</tr>
<tr>
<td>Kebab</td>
<td>Adana</td>
<td>Adana is famous with its Kebab</td>
<td>48.3%</td>
</tr>
<tr>
<td>Scarf</td>
<td>Funeral</td>
<td>Woman use scarf in funerals</td>
<td>48.3%</td>
</tr>
<tr>
<td>Sun</td>
<td>Orange</td>
<td>Both are yellow</td>
<td>48.3%</td>
</tr>
<tr>
<td>Cigarette</td>
<td>Liver</td>
<td>Cigarette is harmful to liver</td>
<td>40.0%</td>
</tr>
<tr>
<td>Poem</td>
<td>Orhan Veli</td>
<td>Famous person who writes poems</td>
<td>40.0%</td>
</tr>
<tr>
<td>Desert</td>
<td>Camel</td>
<td>Camels live in deserts</td>
<td>35.1%</td>
</tr>
<tr>
<td>Rose</td>
<td>Teacher</td>
<td>We give roses to our teachers every teacher’s days.</td>
<td>35.1%</td>
</tr>
<tr>
<td>Beef</td>
<td>Chop</td>
<td>We eat chops of beef</td>
<td>33.1%</td>
</tr>
<tr>
<td>Numbers</td>
<td>Mathematics</td>
<td>We use lots numbers in mathematics</td>
<td>33.1%</td>
</tr>
<tr>
<td>Ball</td>
<td>Wheel</td>
<td>They are very alike</td>
<td>32.6%</td>
</tr>
<tr>
<td>New Year</td>
<td>Santa Claus</td>
<td>Santa Claus and presents comes to our mind</td>
<td>32.6%</td>
</tr>
</tbody>
</table>
4. Conclusion and recommendation

This paper approached the way students use and interact with different metaphors for e-education keywords. First, this research has involved naturalistic processes that are embedded in the same milieu: Cyberspace, the digital world, and the human mind. Secondly, metaphors were considered as expressions or manifestations of the same philosophical foundations.

This paper underlines the importance of metaphors in the process of information communication in eLearning environments. The authors’ recommendations, concluded based on the results of this research, are that the potential of metaphors in e-education should be explored more extensively and more thoroughly in order to support better interconnections among eLearning actors. The use of metaphors creates the setting for a closer and more open communication and increases the sense of social presence in virtual learning spaces, increasing the educational outcomes. Metaphors stand out as a key element that streamlines communication, attracts e-students and increases their involvement in e-experiences. Metaphors build on the sense of community and should be developed to further levels that enhance e-students’ experience and e-teachers’ abilities to communicate the course objectives.

The exploration should focus upon open source as a metaphor for instructional practices – design and delivery, instructional platforms - technologies, and instructional philosophy of e-education. The key questions that deserve attention in the area of instructional practices - design and delivery are:

- Is e-essentially a heuristic experience within the context of a shared repertoire of communal resources (routines, sensibilities, artifacts, vocabulary, styles, etc.) that members have developed over time?
- What does this suggest in regard to the design of e-education?
- What does this suggest in regard to the provision or the delivery of e-education?

Future research shall extend the distribution of the questionnaire in other countries. The authors shall consider both countries whose culture is similar to Turkish and Cypriot culture, and also countries that present a different cultural background.

References


Hara, N, Kling, R. (2003). Learners’ Distress with a Web-based Distance Education Course: An Ethnographic Study of Participants’ Experiences. Turkish Online Journal of Distance Education-TOJDE.4(2). ISSN 1302-6488


