MONITORING BLENDED LEARNING ENVIRONMENTS BASED ON PERFORMANCE DATA

Markus Dahinden
ETH Zurich, Chair of Information Technology and Education
Universitätstrasse 6, 8092 Zurich, Switzerland
markus.dahinden@inf.ethz.ch

Lukas Faessler
ETH Zurich, Chair of Information Technology and Education
Universitätstrasse 6, 8092 Zurich, Switzerland
faessler@inf.ethz.ch

ABSTRACT

A blended learning course with a large number of students is a complex productive system, and measuring its quality is a demanding task. Traditional educational monitoring is based on pre-course and post-course surveys and often measures students' subjective view of the learning environment. As we know, extrinsic motivational elements often have a greater impact on students' opinion of teaching quality than do didactical principles. We therefore define the quality of a course by the amount of knowledge our students have acquired. In consequence, we prefer to use performance data as the basis for learning evaluation. In this paper we present performance data from 284 students which is stored in an educational data warehouse (eDWH). This data has been linked to process data and status data from the same students. Our findings indicate that performance data can be useful in operationalising the quality of a blended learning environment. Based on three concrete examples, we have been able to illustrate the potential of educational monitoring based on performance data. Our monitoring concept not only allows us to verify the validity of our examinations, but also serves to improve our course and to measure this improvement. The results presented here therefore show that the monitoring of a blended learning environment can indeed be improved using performance data.

KEYWORDS

learning evaluation, educational monitoring, educational data warehouse, computer-based assessment, exam validity

1. MEASURING STUDENT PERFORMANCE: A QUESTION OF METHODOLOGY

Designing a blended learning environment is a demanding task, and measuring its quality is difficult. Several evaluation methods exist, the most common being the collection of status data using pre-course and post-course surveys. This is not satisfactory as it mainly summarizes subjective impressions of the students' acceptance of a learning environment. Extrinsic motivational elements such as the entertainment value of the course content often bias student feedback. It is well known that these extrinsic elements do not guarantee improved knowledge, which is in our opinion the most important quality criterion for a learning environment. Therefore, we propose using performance data to measure the outcome and thus the quality of a learning environment.

Establishing a measurement of student performance is not trivial. The result of the final examination, for example, can only be used for educational evaluation if it is valid, i.e., only when this result reflects the cognitive levels of the student with relation to the educational goals. Because it is not possible to determine the real cognitive level of a student, external criteria must be used (Lienert 1998). The choice of the appropriate criteria primarily depends on the educational design of the learning environment. Our tutoring style of instruction provides process data where the tutors' appraisal of students and the students' self-evaluation serve as an independent instrument for measuring student performance.

Over the past 10 years we have developed blended learning courses to teach and assess an introduction to computer science for ETH Zurich natural science students. We teach more than 630 students per year. First we implemented problem-based e-learning methods to make learning independent from time and location, to individualize instruction, and to enhance student time per task (Hinterberger 2007, Faessler 2007). Then we used e-assessment methods to hold and administer examinations electronically. This enabled us to perform formative mock tests with automated corrections and individualized feedback (Dahinden 2010). With these improvements our blended learning environment became more efficient and stable. To observe our progress empirically, we continually analysed data obtained during our courses (Faessler et al. 2005, 2006, Scheuner & Faessler 2010).

In this paper we present and discuss blended learning monitoring procedures based on educational data obtained mainly as a side effect of the grading procedure. These data were collected in an educational data warehouse (eDWH) from students over a period of one semester and linked to process and status data from the same students. To evaluate the data we asked ourselves the following three questions, all of which concern the students' progress in our learning environment:

Question 1: What effect does a mock test have on student performance?
Generally, a mock test implies an additional workload for both teaching staff and students. As we know from student feedback, our mock test is highly regarded and helps students to monitor their progress towards reaching educational objectives. Does the mock test help students to perform better in the final exam? To answer this question, we compared performance data from our mock test with those from the final exam.

Question 2: Can we rely on the validity of our final examination, while grading our students?
The validity of the exam is an important precondition for using performance data as a course's quality criterion. Ideally, an exam measures the students' achievement of the educational goals of a course. In other words, it should reflect the most important learning objectives. Exam validity is difficult to determine, and there is no numerical index for exam validity available. However, external criteria can be used instead (Lienert et al. 1998). To measure the validity of our exam, we wanted to know if students' cognitive capabilities during the semester reflect their performance in the final exam. Students' ratings were tracked over time via a formative assessment conducted by tutors and were then compared to the students' grades in the final exam.

Question 3: Do our formative assessment elements help the students to realistically rate their level of knowledge during the semester?
If we assume that our final exam is valid, we have the possibility to observe interesting psychological factors in students' exam performance in our blended learning environment. For example, it is interesting to observe whether there are differences between self- and external evaluation of student performance during the learning process and final exam. Self-evaluation means that students judge the quality of their own work and knowledge as compared to others. External evaluation is carried out by our assisting tutors and via the final exam.

2. COLLECTING EDUCATIONAL DATA

In this section, we first outline the educational setting of our blended learning course and our measuring instruments, and then we describe how we collected and analyzed our educational data.

2.1 Educational Design

Most of the natural science students at ETH Zurich (630 per year) are introduced to Information and Communication Technology (ICT) through a sequence of 6 learning modules on different topics: exchanging information over the internet, simulation, analyzing multivariate data, managing data with relational databases, and developing macros with VBA (Hinterberger 2010). Figure 1 shows the educational design of our blended learning course. For each module, the students first follow a guided instruction and then work independently through a new, but related problem.

These two steps are implemented according to the 4-step model; a highly effective problem-based instructional method for teaching content while the students work with real data, using suitable application software in a blended learning environment (Faessler 2007). This instructional method ideally supports the combination of problem-based learning with e-learning to accommodate a large number of students independent of time and location. The guided part additionally includes hypertext instructions, which introduce students to basic concepts by processing real-life data with application software (Faessler et al. 2004). Afterwards, during the independent part of the learning sequence, students must independently apply the concepts introduced in a realistic setting. A first formative assessment session takes place at the end of each module, when students have to present their results individually to a tutor of their choice from our tutor pool (formative assessments). At the end of the course, a computer-based final exam is held (summative assessment). Prior to the exam, an optional mock test prepares students for the exam situation.

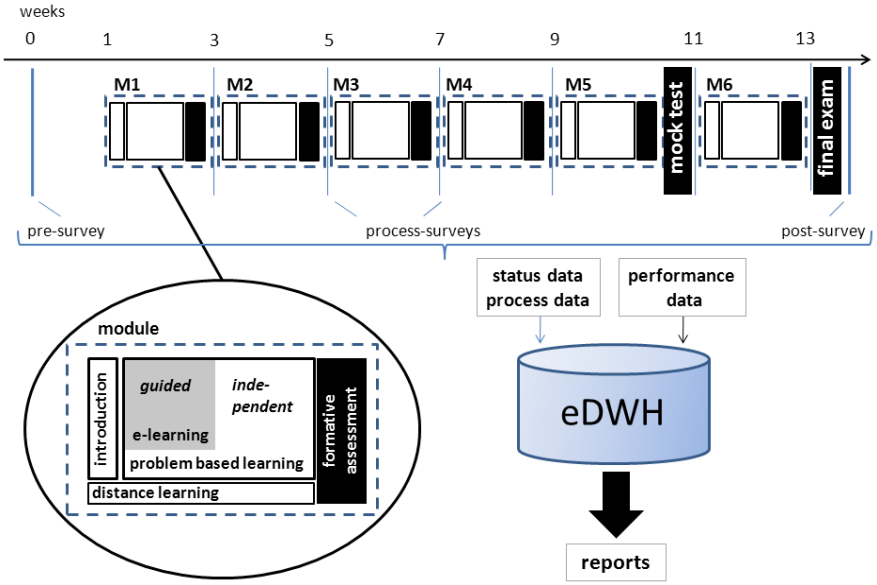

Figure 1. Data collection schedule in our blended learning course. Semester length: 14 weeks. See text for details.

2.2 Our Measuring Instruments

To answer the questions mentioned above, we integrated the following measuring instruments into our educational design (see Figure 1):

Student performance: As an objective measurement of student performance, we used test and exam results. Our exams are computer-based, take 60 minutes and consist of 48 case-oriented questions (Dahinden 2010). A case-oriented exam question involves the presentation of a real-life case study. This case serves as a context in which the students have to apply the concepts learned to answer the given questions. The mock tests comprise half the number of questions of an exam and take half as long.

Tutor rating: To track student performance over time, we used the tutors' formative assessment (6 assessments per student). Tutors were randomly selected to attend the 5- to 10-minute student presentation. Students were graded by tutors according to the following scale: *fail (0)*; *average (1)*; *good (1+)*; *excellent (2)*.

Student self-evaluation: Student self-evaluation took place at eight points (one pre-survey, six process surveys and one post-survey). The process surveys were part of the six learning-modules. As soon as students passed one of their formative assessments, they were asked to complete an online survey delivered automatically by e-mail. It consisted of 6 questions identical for all modules. The survey question focusing on student self-evaluation was: *Compared to my fellow students, my knowledge in this module was…* The

following 5 choices were available: *definitely above-average (5)*; *rather above-average (4)*; *average (3)*; *rather below average (2)*; *definitely below average (1)*. A similar question was asked in the pre- and post-survey.

2.3 Data Collection

At different time points during the learning sequence, three types of data were collected (see Table 1). Status data was collected at the beginning and at the end of the course. It consists mainly of data describing the students in more detail (i.e., branch of study, gender, preliminary knowledge, final grade expectation, etc.). When we talk about process data in the context of educational monitoring, we are referring to data collected at different time points during the semester. Process data mainly contains the students' self-evaluation and tutor ratings of the students' performance at each specific time point. Process data therefore can deliver information about the progress of the students in the learning sequence. Performance data is generated mainly as a side effect of the grading process (mock test and exam).

Table 1. Different educational data collected over one semester (14 weeks).

Data category	Raw data source	Code	Number of aggregated datasets
Status data	Student pre-survey	pre	234
	Student post-survey	post	124
Process data *(status data collected over time*)	Student process surveys	proc	177
	Tutor formative assessments	fAs	284
Performance data	Mock test	t	254
	Final exam	e	276

The data obtained were integrated into an eDWH according to the ETL process (extraction, transformation, load). In this process the various raw data sources were extracted from their source systems, transformed and joined using a mnemonic, allowing us to anonymously trace the answers of each student. Then the data were loaded into the eDWH. Our eDWH is based on a star schema, where the final exam results are included in the centralized fact table and the remaining data are considered as the corresponding dimensions. The eDWH is implemented as a mysql 5.1.4 database and the stored functions are used to encode the raw data dimensions into ordinal scaled datasets in real time.

3. RESULTS

Our eDWH comprises educational data collected from 284 students during Autumn Semester 2010. These data have more than 50 dimensions, which cover measurable facts from one student (e.g. gender, exam score, tutor grading, self-evaluation, etc.). The amount of data allows a wide variety of reports and analyses. In this section, we present some concrete results to illustrate the potential of eDWH-based educational monitoring. The following results refer to the three questions mentioned in the introduction. First, we compare performance data to monitor how students performed on both the mock test and the final exam. Second, we try to measure exam validity by tracking the students' performance data over one semester. Finally, we link performance data with the students' self-evaluation in order to monitor how student self-evaluation evolves over time.

3.1 Comparing Performance Data

Performance data from the optional mock test (n=254, 89% of total students) was linked with the total exam score in Figure 2 using a scatter plot (n=276, 97% of total students). Both dimensions show the percentage of the achieved score compared to the total score. Because the mock test included one question from module 6 (not taught at the time of the mock test; see Figure 1) this question was removed and the score was adjusted accordingly.

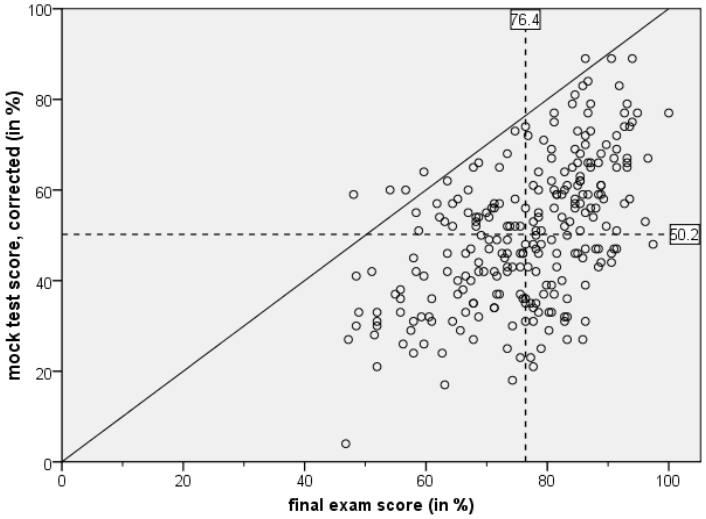

Figure 2. Comparison of the scores of the mock test and the final exam. Each point represents the performance of one student in the mock test (y-axis) and the final exam (x-axis). Both dashed reference lines indicate the mean values of both scores. The diagonal line separates students which performed better in the mock test than in the final exam.

Figure 2 shows a positive correlation between the mock test score and the final exam score. The final exam score is about 26.2% higher than the mock test score. However, there were 5 students (2%) who performed better in the mock test than in the final exam. It is remarkable that 4 out of these 5 students belong to the weakest quartile in the final exam. The remaining 249 students (98%) performed better in the final exam than in the mock test.

3.2 Tracking Student Performance over Time

Tracking student performance over time and linking it to the final exam score can be used to evaluate the validity of the final exam. Figure 3 shows the average of the tutor ratings for all six student presentations.

The students (n=276) are split equally into four quartiles according to their final exam score (EQ). Our data show that the top 25% of the students in the final exam (EQ1) were also assessed as best by the tutors during the semester (m1 to m6). In addition, the weakest students were assessed lower than the rest. It is also evident that module 1 (m1) and module 6 (m6) show the biggest differences in tutor ratings, whereas module 5 shows the smallest difference.

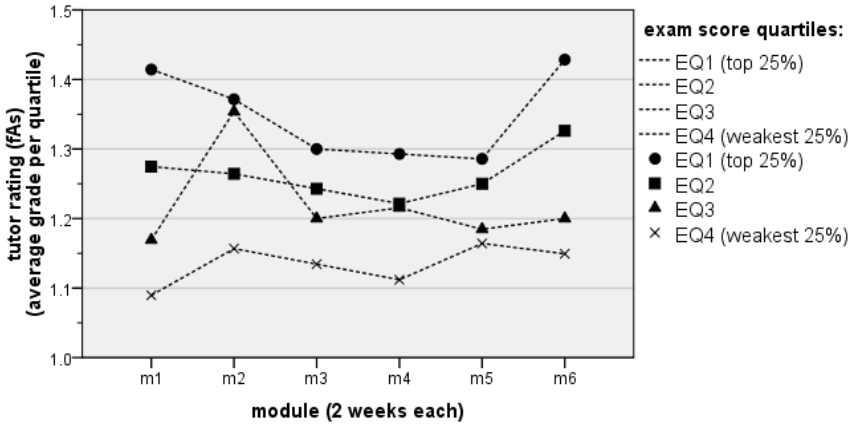

Figure 3. Comparison of the tutor ratings and the final exam result. Students were split into quartiles according to their final exam scores. The top 25% of the students are in EQ1, the weakest in EQ4. The y-axis shows the mean values of the tutor ratings according to the corresponding quartiles (EQ1-EQ4). The x-axis indicates the time variant scale representing the six modules spread over the semester (m1-m6).

Table 2 shows the Spearman's rank correlation (Spearman 1904) coefficients between the final score and nine external criteria. The final score shows significant positive correlations with 8 out of the 9 external criteria. As described above, the tutor rating for module 5 shows the lowest overall significance.

Table 2. Spearman's non-parametric correlations between the final exam grade and nine external criteria: two student final grade estimations (X_ge, 6), six tutor ratings of the students' presentations(fAs_mX), and one student self-evaluation (post_comp). For details see section 2.2.

	pre_ge	fAs_m1	fAs_m2	fAs_m3	fAs_m4	fAs_m5	fAs_m6	post_ge	post_comp
e_score	0.208**	0.320**	0.161**	0.131**	0.169**	0.116	0.242**	0.475**	0.194*
Sig. (2-tailed)	0.006	0.000	0.008	0.031	0.005	0.056	0.000	0.000	0.500
N	273	273	272	272	272	272	271	101	103

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

3.3 Performance Data Linked with Process Data over Time

To give an example of performance data linked with process data, we combined tutor ratings with student self-evaluation over time (Figure 4). The students (n=276) were split into four equal quartiles based on their sub-score in any of the six modules in the final exam (MQ).

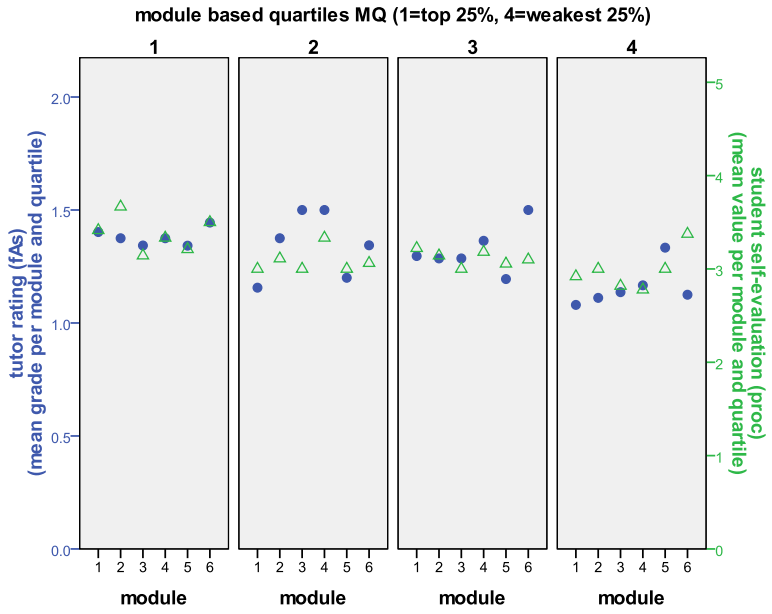

Figure 4. Performance data (tutor ratings) linked to student self-evaluation data. The quartiles are built according to students' sub scores in each module. The top 25% of the students are in the first quartile (MQ1), the weakest 25% in the fourth quartile (MQ4). The mean values of the tutor ratings are shown as dots (●). The mean values of the students' self-evaluation are shown as triangles (△). The scales are aligned according to the values in MQ1.

Figure 4 shows that for the top 25% of the students (in MQ1) the development of the tutor ratings almost follows the development of the average of the students' self-evaluation. This alignment can be used as a baseline for monitoring whether the students in the other quartiles judged themselves higher or lower than the corresponding tutor grading. It can be seen that in general students from MQ2 rated themselves lower compared to their tutors than students in MQ4. Especially for modules 1, 2 and 6 the weakest students rated themselves much higher compared to their tutors. It can also be noted that the students' rating for MQ4, module 6 (mean value=3.4) is almost as high as it is for the best students in MQ1 (mean value=3.6). It can also be seen that the students in MQ3, module 6, rate themselves lower relative than did their tutors' reports.

4. DISCUSSION

With the ability to link performance data with process data and status data, the possibility of educational monitoring arises. To demonstrate the potential of eDWH based monitoring we generated some reports in the result section. Here we discuss these reports with reference to the following three questions:

- Question 1: What effect does a mock test have on student performance?
- Question 2: Can we rely on the validity of our final examination, while grading our students?
- Question 3: Do our formative assessment elements help the students to realistically rate their level of knowledge during the semester?

Q1: What effect does a mock test have on student performance?

The large number of students attending the optional mock test shows us that they are highly motivated to participate. To gain a deeper insight into the students' level of knowledge during the last 2 weeks of the semester, we compared performance data from the mock test with that of the final exam (Figure 2). We found that students performed on average 26.2% better in the final exam than they did in the mock test. The reason for this is unknown and further analysis is needed to find out whether the students' knowledge level really increased or whether they merely worked less for the mock test. However, from the distribution of the raw data, we assume that they participated actively in the mock test. An indication here is the positive correlation between the results of the mock test and the final exam. Another interesting point is the fact that some students showed a remarkable increase in the final exam score compared to the mock test score. This increase could be considered as a "boost-effect" of committed students, but requires more detailed analysis.

Q2: Can we rely on the validity of our final examination, while grading our students?

Due to the fact that in reality it is not possible to determine the real cognitive level of a student, we used process data such as tutor ratings and student self-evaluation data as external criteria for the validity of the exam. Based on the correlation coefficients summarized in Table 2, we can conclude that our exam is valid in terms of the achievement of learning objectives. In other words, the tutor ratings during the semester and the students' self-evaluation reflect the students' performance in the final exam. This effect can also be observed graphically (Figure 3).

Q3: Do our formative assessment elements help the students to properly rate their levels of knowledge?

The focus of this question is to verify whether the students' self-evaluation is different from their external evaluation or not. To answer it, we used tutor ratings as the external evaluation of the students' knowledge. From the graphical representation in Figure 4, we conclude that the tutor ratings for the best students are comparable to the students' self-evaluations. In contrast, the weakest students showed signs of overestimating their own competence, whereas average students showed a tendency to underestimate it. We will use these findings to investigate how we can improve the instruction of our tutors so that they can give feedback to the students which will help the latter to better estimate their level of knowledge.

To sum up, although various technical formative assessment solutions are available (hypertext instructions, computer-based assessment), we consider our tutors to be the most important and the most flexible formative feedback instrument. Due to the fact that our tutors change every semester we must find ways to instruct them efficiently and effectively. We will use monitoring based on the eDWH to determine the effect of the tutor instruction in our blended learning system. Here the data stored in our eDWH will let us both define new improvements, and serve us as a baseline to monitor its effect.

5. CONCLUSIONS & FURTHER WORK

In this paper, we presented some preliminary educational monitoring results based on performance data collected from 274 students during our introductory computer science course. The conclusions we draw are useful in differentiating our understanding of our blended learning environment. As we have seen, performance data play a crucial role in educational monitoring. Its potential for categorising students according to their levels of knowledge is a powerful instrument for improving the understanding of

educational data. In addition, valid performance data can be used as an important quality indicator for the whole blended learning environment: it tells us what and how much the students have effectively learned. We therefore propose using performance data to operationalise the quality of a blended learning environment. With the first round of monitoring presented in this paper, we have set a baseline and can now focus on improving the educational environment. Currently we are using these reports to improve our tutor instruction and learning materials. We will monitor its effects over the next semester.

For the future, we plan to continuously monitor our learning environment and to work toward real-time monitoring. This will allow us to intervene immediately (e.g., if our students' motivation decreases or time expenditure increases). In addition, we will seek answers to open questions such as *What is the effect of a mock test on weak students?* Or, *How is it that tutor ratings are already significant at the beginning of the semester?* Finally, we plan to transfer eDWH-based monitoring to new blended learning systems. This will allow us to check whether any effects of our course have spilled over (e.g., good student effort in mock tests, the effect of weak students' overestimation of themselves, etc).

ACKNOWLEDGEMENTS

We are indebted to all the students at ETH who diligently answered our survey questionnaires and to Hans Hinterberger, Barbara Scheuner and Hans-Joachim Böckenhauer for their critical and helpful comments. The authors would also like to thank Katherine Hahn Halbheer from Lehrentwicklung und -technologie at ETH for proofreading this paper.

REFERENCES

Dahinden, M. (2010). *Anleitung zur Erstellung von problembasierten Pruefungsfragen fuer den ICT-Unterricht*. Technical Report 710, ETH Zürich, Information Technology and Education.

Dahinden, M. & Hinterberger, H. (2010). *Computer-basierte high-stake Leistungskontrolle mit Sioux: Planung, Durchfuehrung und Auswertung einer Basispruefung mit 269 Studierenden*. Technical Report 689, ETH Zürich, Information Technology and Education.

Faessler, L. (2007). *Das 4-Schritte-Modell: Grundlage für ein kompetenzorientiertes e-Learning.* Diss. ETH Zurich, Nr. 17521., Switzerland.

Faessler, L., Hinterberger, H., Dahinden, M. & Wyss, M. (2006). *Evaluating student motivation in constructivistic, problem-based introductory computer science courses.* In T. Reeves & S. Yamashita (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2006*. Chesapeake, USA pp. 1178-1185.

Faessler, L., Hinterberger, H., Bosia, L. & Dahinden, M. (2005). *Assessment as an instrument to evaluate quality of instruction.* In P. Kommers & G. Richards (Eds.*), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005*. Chesapeake, USA, pp. 3555-3562.

Hinterberger, H. (2010). *Making Informatics Work for Everyone - Teaching Computer Competences for the Natural Sciences at ETH*. Technical Report 686, ETH Zürich, Information Technology and Education.

Lienert, G. & Raatz, U. (1998). *Testaufbau und Testanalyse*. 6. Auflage. BELTZ Verlag.

Mietzel, G. (2003). *Pädagogische Psychologie des Lernens und Lehrens*. Göttingen: Hogrefe Verlag.

Scheuner, B. & Faessler, L. (2010). *Log-File Evaluation of a Problem Based e-Learning Unit on Visual Literacy. IADIS International Conference e-Learning 2010, 22.-29. July 2010,* Freiburg Germany, Vol. 2, 171-176.

Spearman, C. (1904) The Proof and Measurement of Association between Two Things, *The American Journal of Psychology*, 15, 72-101. Reprinted International Journal of Epidemiology 2010; 39:1137-50.