ED-MEDIA, 2005, Montreal Canada, pp. 3555 - 3562

Assessment as an Instrument to Evaluate Quality of Instruction

Lukas Faessler, Hans Hinterberger, Laura Bosia and Markus Dahinden
Institute of Computational Science
Swiss Federal Institute of Technology Zurich
CH-8092 Zurich, Switzerland
E-mail: {faessler, hinterbe, bosia, markus.dahinden}@inf.ethz.ch

Abstract: We propose that the assessment of instruction include cognitive competence levels, based on Bloom's taxonomy. Quality control must include the entire instructional process, including assessment. We briefly describe how we organize our instructional process around a student-centered, problem-based learning approach, which allows students to reach the application-oriented competence levels asked for in the exams. The results of exams are useful beyond grading, they can be used to evaluate the quality of the entire instructional process. Some quantitative results illustrate how we have used information provided by exams from the past three years as quality measures and to document the development of our course.

Keywords: assessment, application oriented learning, evaluation, multiple choice test, Bloom's taxonomy, problem-based learning, blended learning.

1. Quality Control Depends on Evaluation Instruments

European Universities are in the process of changing from their traditional diploma programs to a bachelor/master system. The goals and the new structures were defined at the highest level by a commission of the European Communities (Commission of the European communities 2001). The European Ministers of Education point to quality as the basic underlying condition for trust, relevance, mobility, compatibility and attractiveness in the European Higher Education Area. As a consequence, they encourage closer cooperation between mutual recognition and quality assurance networks. They also point to the capability for lifelong learning as another important goal for the restructuring of higher education.

This so-called Bologna Declaration challenges all European Universities to control quality. Mutual agreements to a set of quality standards, however, will not be sufficient. Teachers and course developers also need didactical and technical support in the form of instruments to measure and publish the quality of their instructional units so that comparisons become possible. Such instruments must go beyond traditional evaluation procedures and include a description of the cognitive levels of the teaching goals. This paper describes how we met these challenges.

But how can the quality of instruction and exams be classified? We have chosen as a measure the cognitive levels provided by Bloom's taxonomy of educational goals (Bloom et al. 1956, Anderson et al. 2001), which distinguishes between six classes (K1 - K6): *knowledge*, *understanding*, *application*, *analysis*, *synthesis* and *evaluation*. The taxonomy provides clearly defined teaching goals that describe different levels of a student's competency. The same competencies can also be used to define the level of assessment. Matching the cognitive levels of assessments to those of the instruction is an important aspect of instructional design in general. We describe an example of it in the context of blended learning. As it turns out, mastering higher competence levels is also a necessary prerequisite for lifelong learning.

2. e-Learning can Support but does not Guarantee Quality of Instruction

During the past 4 years, we have developed, used and evaluated so called *E.Tutorials* (previously called application guides) to teach introductory ICT courses (Information and Communication Technology) for natural science students (http://www.et.ethz.ch, see also Faessler 2004, Hinterberger 2004). Each year, 500 to 600 students complete this course, spending a total of 60 to 80 hours on six different topics (including lectures, exercises, review and exam). Students learn skills and concepts to effectively use a computer-based workplace for the following tasks: exchanging information over the internet, presenting information graphically, analyzing multivariate data, modeling and managing data including relational databases, developing macros to adapt user programs to individual requirements. Starting in February 2005 E.Tutorials will also be used in introductory programming courses (Java, Delphi). Because students work in a self-directed fashion it is possible to individualize instruction even within large classes.

E.Tutorials incorporate a problem-based learning (PBL) approach. We have observed that students working with E.Tutorials are more motivated, work more intensively, apply conceptual knowledge more flexibly and independently than in a traditional lecture setting. We also note that over the semester their self confidence increases (Hinterberger, 2004). The fact that students also learned more has so far only been shown with exam results and observation. Both methods, however, do not provide data that allow measurable comparisons. This paper introduces a method that allows us to document changes in instructional quality on the basis of learner's achievements.

Quality control must include exams

An evaluation of instructional units typically includes indicators that measure the students' acceptance of course material, the course's entertainment value and short term satisfaction. The effectiveness of an entire instructional unit, however, which essentially indicates how much and what students actually have learned, has so far often been neglected. One of the major problems is timing: students normally have not completed their learning process at the time when the instructional unit is being evaluated. To get a more complete picture of how learners and instructors can most effectively be supported and which parts of the instructional process should be improved the evaluation must include the entire learning process, including the exam.

The exam can be a useful evaluation instrument to quantify the quality of instruction with respect to course content and teaching goals. But because it is strictly output oriented it does not measure the instructional quality of a course. In the interest of better quality control we would like to combine the two by supplementing traditional evaluation procedures with methods that allow analyses of a learner's output.

The role of assessments

Assessments are in general limited to providing data for a grading process which typically takes place at the end of an instructional unit. Grading often fails to include the instructional process as a whole. Instructors want to give a good and interesting course, but they also have to separate successful students from unsuccessful ones. To be of any value, the selection must take place at a cognitive level that corresponds to the cognitive level which underlies instruction. Often exams are perceived as difficult not because the students are ill prepared but because the wrong competencies are tested. Sometimes, however, instead of correcting such a mismatch, instructors use the difficulty of their exams as an indicator for the quality of their instruction. The real problem, however, is that students are not able to apply their theoretical knowledge because it remains inactive.

Students are primarily interested in passing exams with a reasonable effort and regard them as a chance to show that they learned something interesting and useful. Both students and instructors are interested in successful exams, administered at a high but attainable level.

Wilson (1994) proposed a so called 'Redesigning process' for education and suggested that quality can improve if all aspects of education are accounted for by treating the whole process as a single "product", much in the way engineers approach their work. In consequence assessment must be included in the educational design process. Teaching goals must be defined not only for instructional units but also for exams and the two must be compatible.

This approach can be compared with how a conductor and orchestra work together towards a single event, the concert. This also means, however, that the competencies and knowledge of students have to be developed before they can be tested with an exam. From this point of view, an exam should never assess more than has been taught in the corresponding lecture. It can, however, be used not only as a means for selection but also as an evaluation instrument to provide information about what competencies the students actually have developed during instruction.

Exams can take on a new role as an evaluation instrument for education, be used to improve the instructional units and also to document how instruction evolves over time.

3. Quality Assurance of Instruction Based on Cognitive Competence Levels

In traditional instructor-centered courses (predominantly lecture-based), learners typically reach levels K1 and K2 (knowledge and understanding). Levels K3 and higher (application, analysis and synthesis), henceforth called K3+, are difficult to reach because students often do not have the chance to become active enough and therefore work only superficially through a subject. They do not learn to apply their newly acquired knowledge, because they only collect information instead of using it to solve concrete problems. To be able to apply their knowledge in an exam, students develop strategies of their own and spend much additional effort reading, memorizing and thinking through the subjects all over again. During this process students need additional support, otherwise they will still not learn how to apply their knowledge, will still have difficulties with their exams, and might get frustrated or even give up. This problem is often solved by reducing exams to levels K1 and K2. Frequently the questions requiring levels K3+ are eliminated when the exams are corrected in order to achieve "acceptable" grades.

The problem with many exams is not that they are too difficult but that the students did not have the chance to acquire the right competencies.

Instruction leading to competence levels K3+

To avoid this competency mismatch, we utilize PBL in conjunction with constructivist methods applied in a self-directed way. This combination is necessary because PBL by itself does not necessarily include the detail of instruction necessary for novices, whereas a constructivist method alone will remain just that, a method. In this way we are able to provide courses that are optimized to support K3+ learning. The vehicle for this are the *E.Tutorials* and the subsequent oral *certifying test* (Fig. 1, Hinterberger 2004).

Students learn while working through a demanding and realistic problem. The success of PBL can be explained with results from brain researchers who found that rules or theories can only be embedded in the brain if they are actively used while the learner engages with a real system (Spitzer 2002). During this engagement, a learner's brain demonstrably generates the necessary rules by itself. This phenomenon is also the basis for the constructivist approach: everybody constructs their knowledge by himself or herself.

If one regards learning as a process in a complex system, then the possibilities for control from the outside are very limited (Malik 2003). Consequently we provide our students with much flexibility regarding time, place and duration when working through the course materials under their own control (self directed).

To support students at the beginning of their construction process, and therefore prevent frustrating experiences, learners must initially be given exact instructions on how to construct and solve the problem they are presented with. This is a necessary prerequisite for any learner to become increasingly independent and to gain self confidence.

A potential side effect of PBL is that learners understand the problem, but might have difficulties in disengaging from it and reaching a level that allows generalizations. To verify that the students achieve this level, we introduced oral presentations to let the them summarize in their own words what they have done. We also ask them test questions at the end of each instructional unit. Classroom lectures can also support this process by showing "the big picture" into which the work done with PBL fits.

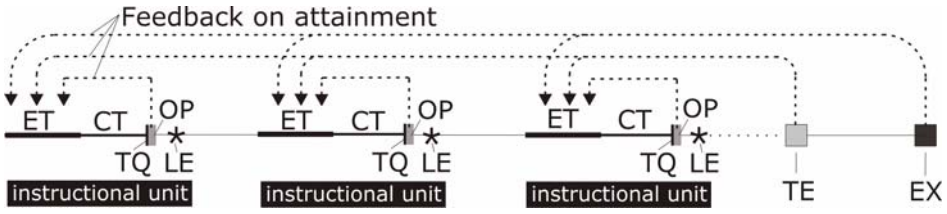

Figure 1: K3+ based courses consist of several instructional units and associated feedback on attainment. The E.Tutorial (ET) gives instruction that lead to competencies, which can then be applied independently in an appropriate certifying task (CT). Test questions (TQ) and an oral presentation (OP) complete one learning unit and prepare for K3+ based testing; these are also the main components of the feedback loop. Personal contact between students and instructor during a lecture (LE) also supports this process. A trial exam (TE) prepares for the final exam (EX) by simulating the latter in content and form.

Assessment at levels K3+

Contrary to widespread opinion, competencies at levels K3+ can be assessed with multiple choice tests (Bloch 1999). Much like PBL, this is accomplished through solving a given realistic problem (e.g. interpreting a small piece of program code) and selecting from a choice of solutions the correct ones (k out of n). Students show their competencies by applying their knowledge in the given task. These types of application-oriented multiple choice questions are time-consuming and demanding to produce, but they are quickly corrected and the results can be readily analyzed. They are also quite revealing, both with respect to the assessment of the students as well as the quality of the test.

The sequence "Instruction → Problem Based Task → Oral Presentation" alone is not enough to prepare students so that they can readily apply the concepts flexibly in an exam. It is characteristic of cognitive levels K3+ that students need to acquire a thinking pattern which can be used to solve a given problem instead of simply reproducing knowledge. The best musicians need rehearsals to prepare for a good concert. To prepare our students for K3+ levels we provide them with training questions after each instructional unit and have them attend a trial exam at the end of the course. Both let the students know what to expect from the application-oriented final exam in general and how to deal with problem-based questions in particular. In order to solve the problems of the K3+ exams, students cannot resort to memorizing facts, they must be familiar with the underlying concepts, independent of the details or the topic of the exam questions. K3+ level instruction and assessment therefore provide students with additional motivation for “deep” learning.

Perhaps the most important condition for successful K3+ assessment is that instruction and assessment take place at the *same* cognitive level and that learners have the possibility of training the mode of thinking that will be expected of them. The mental effort that has to be expended between the first construction of a concept (while working through a problem) and its flexible application with the new problem of an assessment must not be underestimated. Oral presentations and trial exams help to further develop the thinking patterns required; the output from both can be used, together with the results of the final exam, to evaluate the quality of instructional processes (see Fig. 1).

The next section documents the development of one of our courses towards K3+ assessment and illustrates the quantitative improvement of the students' competencies.

4. Quantitative Results and Discussion

After three years of K3+ instruction we decided to use the exam as an instrument to quantify the quality of our instructional units. To this end we used the results of six final exams from the past three years (Fig. 2). The exams had 33 to 42 multiple choice questions with K-levels 1 to 3 and higher. For each exam we generated new questions. The spring exams contain many students who have to repeat it, making comparisons difficult. We therefore only

analyzed the fall exams more closely (2001/II, 2002/II, 2003/II) which were written by 109 to 276 students (Fig. 3). The analysis of the performance with K3-questions was further restricted to students in pharmacology (46 to 77 students, Fig. 4).

Results

The following three points summarize our experiences:

- During the past three years the number of exam questions at the levels K1 and K2 were reduced in favor of levels K3+ (Fig. 2).
- The distribution of the fraction of correctly answered questions became broader when the proportion of K3+ questions was increased (Fig. 3).
- The number of correctly answered K3+ questions increased significantly during the last 3 years (Fig. 4).

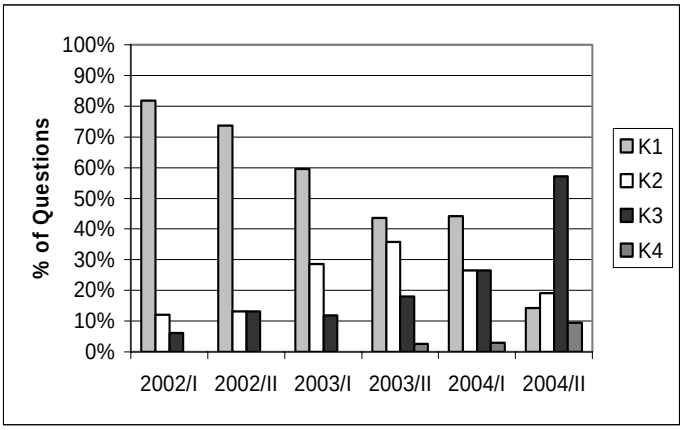

Figure 2: Structure of the exams with respect to the number (N) of questions having a particular cognitive level, for the years 2002, 2003 and 2004. In each year two different exams were held (one in the fall, another in the spring). K1: knowledge, K2: understanding, K3: application, K4: analysis. N: 33 to 42 questions.

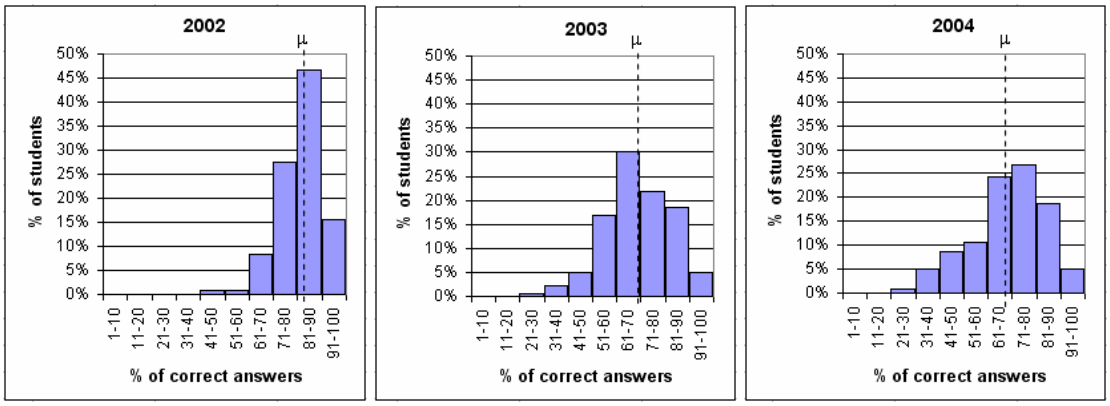

Figure 3: Distribution of the fraction of correctly answered questions for the exams in 2002, 2003 and 2004 written by N students. **2002:** N=109, μ=82.8% σ^2=3.55. **S2003:** N=220, μ=69.5%, σ^2=0.221. **2004:** N=276, μ=68.4, σ^2=0.231.

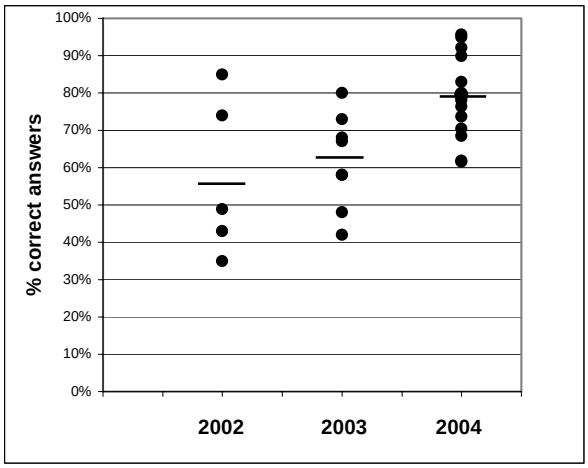

Figure 4: Distribution of the fraction of correctly answered K3-questions for the exams in 2002, 2003 and 2004. Each dot represents a particular exam question. The number of K3-questions was 5 in 2002, 8 in 2003 and 14 in 2004. The proportion of correct answers increased significantly from one year to the next ($p < 0.01$ in Mann-Whitney test). (**2002:** N=49, μ=57%. **2003:** N=94, μ=62%. **2004:** N=77, μ=79%).

Discussion

We briefly discuss our findings based on the three points listed at the beginning of this section.

During the past three years the number of exam questions at the levels K1 and K2 were reduced in favor of levels K3+.

This development illustrates our efforts to increasingly orient our instruction towards application based learning and assessment. In 2002 our instructional units incorporated PBL, but the assessment did not. If at that time we had included in our exams the K3+ questions we ask today, students would not have been competent enough to solve these problems, simply because they lacked PBL practice. On the other hand, the K1 and K2 dominated exam from 2002 would not be satisfactory today because students would not be able to demonstrate their K3+ competencies.

We had to reduce the number of questions slightly, because application-oriented questions need more time to process than questions which only test knowledge and understanding.

The distribution of the fraction of correctly answered questions became broader when the proportion of K3+ questions was increased.

In the K1 and K2 based exam in 2002, the students' performance is characterized by a narrowly dispersed distribution, skewed towards the higher grades. Almost 90% of the students had 70% and more of the answers correct and only 1% of the students produced less than 50% correct answers. This complicates the grading process because only few correct answers separate one grade from the next. In consequence, the selection point between pass and fail has to be chosen in a steep part of the curve of the distribution, which can be uncomfortable. The distribution of correctly answered questions two years later is more broadly dispersed, with a stronger leader group than in previous years and a slower decrease of the lower grades. Over 50% of the students answered 70% and more of the questions correctly and 25% of the students answered less than 50% of the questions correctly. Under these circumstances the choice between pass and fail affects fewer students when the boundary line is moved in either direction.

The number of correctly answered K3+ questions increased significantly over the last 3 years.

Figure 4 indicates that in 2004 students were better able to handle the K3+ questions. They were significantly ($p < 0.01$ in Mann-Whitney test) better than the students in 2002. Mistakes are now spread more evenly over the entire exam than they were two years before, when mistakes were primarily in the few K3 questions. These developments show up in the low grades obtained for the K3 questions while at the same time the grades for the rest of the exam were rather high (Fig. 3).

The distribution of correctly answered K3-questions improved primarily because students had the chance to better prepare themselves for an exam that tested them better at the competency level they acquired during instruction. In 2002 students attended a more traditional course with lectures and a textbook, accompanied with e-learning based exercises. At that time they predominantly read and memorized the concepts out of the textbook, as they would do in any traditional lecture, because they were not made aware of the requirements for K3+ assessment. As a result they were competent to understand and reproduce that knowledge (K1 and K2) but not to apply it (K3+).

Beginning in 2004, our attention shifted increasingly towards PBL, putting the student at the center. Students now know that they not only have to reproduce factual knowledge and concepts, but that they also have to apply them in the exam. We provide additional support by letting them apply the concepts they learned in mock test situations with feedback (Fig. 1). Without this, they would find it difficult to generalize beyond the problem which they learned to understand and which they used to generate the rules that they should apply independently.

During the exam in 2004, students could finally use more of what they actually learned and consequently demonstrate their competencies. This resulted not only in better K3+ performances, it also proved to the students themselves that they could perform at a higher cognitive level. This was confirmed when we asked our students after the exam whether the exam was such that they could show what they learned in the course. 85% of the students said that this was the case. This and many personal comments indicated that the majority of the students found the exam to be fair. In the same survey only 14% of the students stated that they used the lecture notes to prepare for the exam, whereas 87% used the E.Tutorial. We assume that if we had asked this question two years before, the percentages would have been reversed.

5. Conclusions and Further Work

Our quantitative analyses support our initial conjecture that efforts to raise the quality of instruction must coordinate feedback on attainment and assessments at comparable cognitive levels. Classifying exams based on cognitive levels is particularly helpful when qualitative changes must be documented. This should not come as a surprise because education can be regarded as a complex system, where instruction and assessment are part of a feedback loop. Consequently, assessment on a high cognitive level needs as a prerequisite individualized instruction on the same cognitive level, combined with PBL that leads to the thinking patterns expected in the exam. To this end we now build our instruction around the problem-based E.Tutorials.

From an exam, one can extract more information than what is needed for selection; in particular its results tell us what the students effectively learned. Furthermore, the composition of an exam with respect to the cognitive level of its questions (Fig. 4) can be taken as an evaluation instrument to measure the quality of instruction which can be used by instructors and administrators alike. The results can be used not only to improve the instructional units for future courses but also to document any changes that took place historically. Last but not least it will provide the basis for administering good and fair exams.

K3+ instruction and K3+ assessment applied individually and in isolation, however, will not be sufficient to develop K3+ competencies of our students. As illustrated in Fig. 1, instruction and assessment have to be regarded as two aspects of a complex educational process. The primary aim is to teach knowledge and skills in such a way that students will be able to apply it flexibly. The Part of the feedback loop mentioned above is to provide the possibility for repeated practice. It is this kind of skill, rather than the detailed knowledge, that will provide a basis to support lifelong learning.

One of the major consequences of our experience is the realization that e-learning tools by themselves are not enough. In the case of PBL, the tools must be integrated very carefully and sensitively. The effort to achieve a successful integration that supports higher cognitive levels must not be underestimated.

Outlook

We plan to extend our integrated learning method to other subjects and other institutions, in order to support self-directed sustained learning. Additionally we are working on an assessment data base that allows us to store our K3+ exam questions, generate statistics about their use in exams and carry out analyses to evaluate the quality of our instruction.

References

Anderson, L., Krathwohl, D. (2001). *A Taxonomy for Learning, Teaching, and Assessing.*NY: Addison Wesly Longman.

Bloch, R., Hofer, D., Krebs, R. (1999). *Kompetent pruefen. Handbuch zur Planung, Durchfuehrung und Auswertung von Facharztpruefungen.* Bern, Wien: Selbstverlag.

Bloom, B.S. (Ed). (1956). *Taxonomy of Educational Objectives Handbook 1: Cognitive Domain.* New York: Longman, Green & Co.

Commission of the European communities (2001). *Making a European area of lifelong learning a reality*. http://europa.eu.int/comm/education/policies/lll/life/communication/com_en.pdf.

Faessler, L., Hinterberger, H., & Bauer-Messmer, B. (2004). The Application Guide: Each Student His Own Tutor. *World Conference on Educational Multimedia, Hypermedia and Telecommunications 2004*(1), 2371-2373.

H. Hinterberger, L. Faessler, B. Bauer-Messmer: From Hybrid Courses to Blended Learning: A Case Study. *International Conference on New Educational Environments (ICNEE), 2004*, University of Neuchatel, Switzerland.

Malik, F. (2003). *Strategie des Managements komplexer Systeme. Ein Beitrag zur Management-Kybernetik evolutionaerer Systeme.* Bern, Switzerland: Haupt Verlag.

Spitzer, M. (1999*). The Mind Within the Net: Models of Learning, Thinking, and Acting.* Cambridge, Massachusetts: MIT Press.

Wilson, K. G., Daviss, B. (1996). *Redesigning education.* New York: Teacher College Press.

Acknowledgements

This work is based on research supported by Fonds Filep of ETH Zurich. For more information about the project, see http://www.ethworld.ethz.ch/projects/details_EN?project_id=144. We thank Sarah Shephard from the didactic center of ETH for her helpful comments.